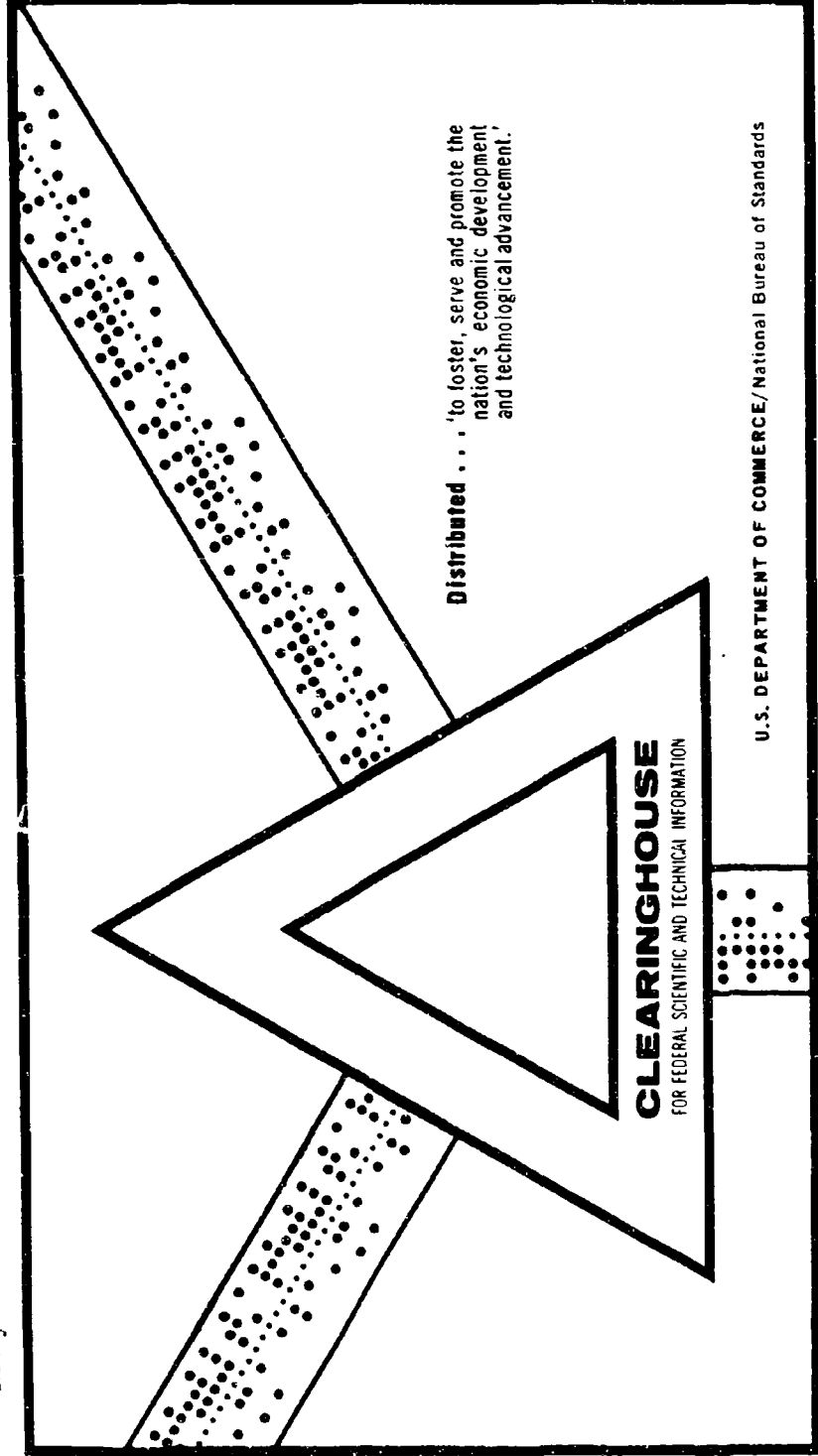# ON DIRECT METHODS FOR SOLVING SYMMETRIC SYSTEMS OF LINEAR EQUATIONS

James Raymond Bunch

California University
Berkeley, California

May 1969

Distributed . . . 'to foster, serve and promote the nation's economic development and technological advancement.'

CLEARINGHOUSE
FOR FEDERAL SCIENTIFIC AND TECHNICAL INFORMATION

U.S. DEPARTMENT OF COMMERCE/National Bureau of Standards

On Direct Methods for Solving Symmetric
Systems of Linear Equations

James R. Bunch

May 1969

Technical Report No. 33

Computer Center
University of California
Berkeley

On Direct Methods for Solving Symmetric Systems

of Linear Equations

by

James Raymond Bunch

Abstract

There has been no stable direct method for solving symmetric
systems of linear equations which takes advantage of the symmetry. If
the system is also positive definite, then fast, stable direct methods
(e.g., Cholesky and symmetric Gaussian elimination) exist which preserve
the symmetry. These methods are unstable for symmetric indefinite sys-
tems. Such systems often occur in the calculation of eigenvectors.
Gaussian elimination with partial or complete pivoting is currently
recommended for solving symmetric indefinite systems, and here symmetry
is lost.

We present a generalization of symmetric Gaussian elimination,
called the diagonal pivoting method, in which pivots of order two as
well as one are allowed in the decomposition. We show that the diagonal
pivoting method for symmetric indefinite matrices takes advantage of
symmetry so that only $\frac{1}{6} n^3$ multiplications, at most $\frac{1}{3} n^3$ additions,
and $\frac{1}{2} n^2$ storage locations are required to solve $A x = b$, where A
is a non-singular symmetric matrix of order n . Furthermore, we show
that the method is nearly as stable as Gaussian elimination with complete
pivoting, while requiring only half the number of operations and half
the storage.

We include a listing of an Algol procedure for the diagonal pivoting method, which is applicable both to symmetric definite and indefinite systems.

We discuss the problem of symmetric band matrices and present an algorithm only for the tridiagonal case. Further, we discuss the problem of equilibrating symmetric matrices while preserving symmetry and we present a simple algorithm (and Algol procedure) for accomplishing this.

Table of Contents

Acknowledgment

This dissertation was prepared under the direction of
Professor Beresford N. Parlett, Chairman, Department of Computer
Science, Berkeley. I am most grateful to Professor Parlett for
his being willing to discuss innumerable contentions and for the
generous contribution of his time that he made.

I wish to thank Professor William Kahan for several helpful
conversations, and I wish to thank the dissertation committee,
Professor Parlett, Professor C. Keith Miller, Department of Mathe-
matices, and Professor Hugh D. McNiven, Department of Civil Engineering,
for reading the manuscript.

I also thank Mrs. Geri Stephen for great patience and exceptional
skill in typing the manuscript.

Chapter 1 : Introduction

1.1   Symmetric Systems of Linear Equations

Let us consider direct methods for solving the system of linear algebraic equations, $A x = b$, where A is symmetric.

If A is also positive definite, then Cholesky's method (§2.7) and the $L D L^t$ method (§2.6) are fast, stable, and preserve symmetry.

If A is symmetric but indefinite (neither positive definite nor negative definite), Cholesky's method and the $L D L^t$ method are unstable and can produce very inaccurate results (§2.8).

At the present time, if A is symmetric indefinite, Gaussian elimination with partial or complete pivoting is recommended for solving the system (Fox, p. 80, 185), and thus the symmetry of A is ot no advantage.

Is there an algorithm for the symmetric indefinite case wh'ch is stable, is faster than Gaussian elimination, and can take advantage of the symmetry?

1.2   Our Contribution

We discuss the problem is Chapter 2 and review previous efforts in Chapter 3. In Chapters 4-6 we present a method, called diagonal pivoting, which fulfills the above requirements when restricted to equilibrated matrices. In Chapter 7 we present a method for equilibrating symmetric matrices in a very simple manner. A variation of the diagonal pivoting method is presented in Chapter 8; this method is applicable to unequili-

brated matrices and it fulfills the above-mentioned requirements. In
Chapter 9 we show that the diagonal pivoting method is almost as fast
as Cholesky or $L\,D\,L^t$. In Chapter 10 we perform a backwards error
analysis. In Chapters 11-12 we show that the method is essentially as
stable as Gaussian elimination with complete pivoting (in the sense of
Wilkinson's analysis for Gaussian elimination with completely pivoting,
Wilkinson (1961)). In Chapter 13 we show that iterative improvement is
as applicable here as it is for Gaussian elimination. In Chapter 14
we discuss the problem of symmetric band matrices.

All the results proved are applicable to complex systems where A
is Hermitian.

1.3   Summary

Let A be an $n \times n$ matrix with $\max_{ij} |A_{ij}| = 1$.

We want to solve $A\,x = b$.

Let $\sim C_k\,N^k$ denote $C_k\,N^k + \sum_{i=0}^{k-1} C_i\,N^i$, where the C_i , $0 \leq i \leq k$,

are constants independent of n .

Let G.E. denote Gaussian elimination.

The situation is summarized in the following table:

Method	Restrictions on A	Number of Multiplications	Number of Additions	Storage	Bound on Element Growth (Stability)
G.E. with complete pivoting	$\det A \neq 0$	$\sim \frac{1}{3} n^3$	$\sim \frac{2}{3} n^3$	$\sim n^2$	$\sqrt{n}\ f(n)$
G.E. with partial pivoting	$\det A \neq 0$	$\sim \frac{1}{3} n^3$	$\sim \frac{1}{3} n^3$	$\sim n^2$	2^n
Cholesky Method	symmetric positive definite	$\sim \frac{1}{6} n^3$	$\sim \frac{1}{6} n^3$	$\sim \frac{1}{2} n^2$	1
Diagonal Pivoting	symmetric, $\det A \neq 0$	$\sim \frac{1}{6} n^3$	$\geq \sim \frac{1}{4} n^3,$ $\leq \sim \frac{1}{3} n^3$	$\sim \frac{1}{2} n^2$	$\sqrt{n}\ f(n) \times c(\alpha)\ h(n,\alpha)$

Here $0 < \alpha < 1$, $f(n) = \left\{ \prod\limits_{k=2}^{n} k^{\frac{1}{k-1}} \right\}^{1/2}$, $h(n,\alpha)$ is a function

dependent on the pivotal strategy, and $c(\alpha)$ and $h(n,\alpha)$ are defined in

§§11.6-7. In Chapter 12 we show that $c(\alpha)h(n,\alpha) < 3.07(n-1)^{0.446} < 3\sqrt{n}$

for $\alpha = (1 + \sqrt{17})/8 = \alpha_o$.

1.4 Origin of Symmetric Indefinite Systems

The problem of indefinite systems of linear equations is sometimes

dismissed as academic by the claim that physical problems always generate

positive definite systems of linear equations. However, the numerical

solution of natural problems often gives rise to situations which do not

have a physical origin. We give two related examples.

In the Rayleigh Quotient Iteration for finding eigenvalues of a

positive definite matrix A (Wilkinson (1965), p. 172, and particularly

p. 629) we need to solve the systems $(A - r_i I) x_{i+1} = x_i$, where

$r_i = x_i^* A x_i / x_i^* x_i$. Here $A - r_i I$ cannot be definite because

r_i lies between the extreme eigenvalues.

In the inverse iteration method for finding the eigenvector corresponding to an approximation λ to an intermediate eigenvalue of a positive definite matrix A, we need to solve $(A - \lambda I) v_{i+1} = u_i$, $u_{i+1} = v_{i+1} / \max (v_{i+1})$. $A - \lambda I$ can be indefinite even when A is positive definite (Wilkinson (1965), pp. 618-635).

Chapter 2 : <u>Presentation of the Problem</u>

2.1   <u>Introduction</u>

The speed and storage capacities of current digital computers
allow us to solve large systems of linear equations by direct methods.
Here we shall consider direct methods for solving a system of linear
equations, $A x = b$, where A is symmetric and $\det A \neq 0$.

2.2   <u>General Problem</u> : <u>Exact Arithmetic</u>

First let us consider the solution, in exact arithmetic, of
$A x = b$, where A is general and $\det A \neq 0$. We know that Gaussian
elimination will give the solution provided that whenever a zero appears
in the leading diagonal position we interchange that row with a lower
row with non-zero leading element (such a row will exist since
$\det A \neq 0$), e.g. if $A_{11} = 0$ and if j is the least integer for which
$A_{j1} \neq 0$, then we interchange rows 1 and j . Or, equivalently, there
exists a permutation matrix P such that Gaussian elimination without
interchanges applied to $P A$ will give us the solution.

Since we could also do the same with columns instead of with rows,
there exists a permutation matrix Q such that Gaussian elimination
without interchanges applied to $A Q$ will give us the solution.

In matrix notation the Gaussian elimination algorithm factors A
into $A = L U$, where L is unit lower triangular, U is upper triangular,
and L and U are unique when they exist (Wilkinson (1965), p. 204).
Thus in exact arithmetic there exist permutation matrices P and Q

such that $PA = L_1 U_1$ and $AQ = L_2 U_2$, provided only that
$\det A \neq 0$.

2.3 General Problem : Failure of Previous Attempts in Finite Precision Arithmetic

In finite precision arithmetic (Wilkinson, 1963) the above algorithm can fail if we interchange only under the condition that the element in the leading diagonal position be zero.

Let $A = \begin{bmatrix} \epsilon & 1 \\ 1 & \eta \end{bmatrix}$ and $b = \begin{bmatrix} 1/\epsilon \\ 0 \end{bmatrix}$.

Then $A = LU = \begin{bmatrix} 1 & 0 \\ 1/\epsilon & 1 \end{bmatrix} \begin{bmatrix} \epsilon & 1 \\ 0 & \eta - 1/\epsilon \end{bmatrix}$. However, if ϵ and η are

small enough, then in finite precision arithmetic the operation $\eta - 1/\epsilon$

yields $-1/\epsilon$.

Let U_c and x_c be the matrix and vector of the values of U

and x, respectively, computed in finite precision.

Then $U_c = \begin{bmatrix} \epsilon & 1 \\ 0 & -1/\epsilon \end{bmatrix}$ and $a = L^{-1} b = \begin{bmatrix} 1/\epsilon \\ -1/\epsilon^2 \end{bmatrix}$.

So $x = U^{-1} a = \begin{bmatrix} \dfrac{-\eta}{\epsilon(1-\eta\epsilon)} \\ \dfrac{1}{\epsilon(1-\eta\epsilon)} \end{bmatrix} \doteqdot \begin{bmatrix} -\eta/\epsilon \\ 1/\epsilon \end{bmatrix}$, but $x_c = U_c^{-1} a = \begin{bmatrix} 0 \\ 1/\epsilon \end{bmatrix}$

If $\eta = c \epsilon$, then $x \doteqdot \begin{bmatrix} -c \\ 1/\epsilon \end{bmatrix}$, and the error in the first component

of the computed solution is $|c|$.

2.4   Condition of a Matrix

Wilkinson (1965, pp. 189-191) shows that the relative error in the solution of a system of linear equations is bounded in proportion to the condition number of A, $\kappa(A) = \|A\| \, \|A\|^{-1} \geq 1$, i.e., if $e = x - x_c$ then $\|e\| / \|x\| \leq \rho(A) \, \kappa(A)$. We should not expect a small error if $\kappa(A)$ is large.

Here A has a very satisfactory condition number. $A^{-1} = \dfrac{1}{\varepsilon\eta-1} \begin{bmatrix} \eta & -1 \\ -1 & \varepsilon \end{bmatrix}$.

Using the one-norm, $\|A\| = \max_j \sum_{i=1}^{n} |A_{ij}|$, we have $\kappa(A) = \|A\|_1 \, \|A^{-1}\|_1 = \dfrac{[1 + \max(|\varepsilon|, |\eta|)]^2}{|1 - \eta\varepsilon|} \doteq 1$.

The computer replaces $\varepsilon\eta - 1$ by -1, and the computed inverse is $(A^{-1})_c = \begin{bmatrix} -\eta & 1 \\ 1 & -\varepsilon \end{bmatrix}$. Then $(A^{-1})_c \, b = \begin{bmatrix} -\eta/\varepsilon \\ 1/\varepsilon \end{bmatrix}$. Thus the trouble lies in the Gaussian elimination algorithm, not in the matrix A.

2.5   General Case : Stable Direct Methods

(a) Direct Inversion

Direct inversion (the formation of A^{-1}) of a system requires $\sim n^3$ each of multiplications and additions (Fox, pp. 177-179). Gaussian elimination, however, requires only $\sim \frac{1}{3} n^3$ each of multiplications and additions. Thus we would prefer to use Gaussian elimination if we could obtain a satisfactory solution.

(b) Stability

Let us attempt to solve $A x = b$, where A is $n \times n$ and $\det A \neq 0$. If x_c is the solution we obtain from the computer, we may consider x_c to be the exact solution of the system $(A + E) y = b$. We might say that the algorithm we use is stable if the elements of E are small in comparison to the corresponding elements of A. Actually the term is more often used when $\|E\| / \|A\|$ is small. (Here $\| \|$ is any norm.)

(c) Stability for Gaussian Elimination

Wilkinson (1960) showed that for Gaussian elimination we have:

$$\left| E_{ij} \right| \leq 2.01 \max(i-1,j) 2^{-t} \max_{r,s} \left| A_{rs}^{(n-k+1)} \right| , \text{ where } t \text{ is the number}$$

of binary digits in the machine, $k = \min(i,j)$, and $A^{(n-k+1)}$ is the reduced matrix of order $n-k+1$ in the elimination process.

The important lesson from the above is that we must be interested in keeping the elements in the reduced matrices small. There are two well-known strategies for choosing permutation matrices P and Q such that Gaussian elimination without interchanges applied to $P A Q$ will provide sufficiently small element growth in the reduced matrices.

(d) Complete Pivoting

The first strategy, called complete pivoting, requires that we bring the largest element in the reduced matrix into the leading diagonal position. This strategy is called complete since we search the entire reduced matrix. Wilkinson showed that this complete strategy gives

$$\max_{i,j} \left| A_{ij}^{(n-k+1)} \right| \le \sqrt{k}\, f(k) \max_{i,j} \left| A_{ij} \right| \quad , \text{ where } \quad f(k)^2 = \prod_{r=2}^{k} r^{\frac{1}{r-1}} \quad .$$

In words, the elements in the reduced matrices can never become too large; so this strategy is never bad. It is conjectured that the true bound is $\max\limits_{i,j} \left| A_{ij}^{(n-k+1)} \right| \le k$ where A is real (§12.4).

Equivalently, the above says that there exist permutation matrices P and Q such that Gaussian elimination without interchanges applied to $P\,A\,Q$ gives $\max\limits_{i,j} |E_{ij}| \le 2.01\, n^{3/2}\, f(n)\, 2^{-t} \max\limits_{i,j} |A_{ij}|$.

(e) Partial Pivoting

The second strategy, called partial pivoting, requires that we bring the largest element in the first column of the reduced matrix into the leading diagonal position. This strategy is called partial since we search only a part of the reduced matrix. This is equivalent to the application of Gaussian elimination without interchanges to $P\,A$, where P is a permutation matrix. Here $\max\limits_{i,j} \left| A_{ij}^{(n-k+1)} \right| \le 2^k$.

This bound is sharp since $A = \begin{bmatrix} 1 & 0 & 0 & & 0 & 1 \\ -1 & 1 & 0 & & 0 & 1 \\ -1 & -1 & 1 & & 0 & 1 \\ & . & & & & . \\ & . & & & & . \\ -1 & -1 & -1 & \ldots\ldots & -1 & 1 \end{bmatrix}$, where A is

$n \times n$, has $A_{nn}^{(n)} = 2^n$. Thus $\max\limits_{i,j} |E_{ij}| \le 2.01\, n\, 2^{-t}\, 2^n \max\limits_{i,j} |A_{ij}|$,

and this is very weak when $n > t$.

Correspondingly, we could use a partial pivoting strategy in which we bring the largest element in the first row of the reduced matrix into

the leading diagonal position. Thus there exists a permutation matrix Q such that Gaussian elimination without interchanges applied to $A\,Q$ has an error matrix E with $\max\limits_{i,j} |E_{ij}| \leq 2.01\ n\ 2^{-t}\ 2^n\ \max\limits_{i,j} |A_{ij}|$.

(f) The Error Matrix

We see that the error matrix E is dependent on the decomposition L, U , the matrix A , the right hand side b , and the permutations P and Q by which we can pre- and post-multiply A , i.e. $E = E(L, U, A, b, P, Q)$, where $L\,U = P\,A\,Q$. (For the partial pivoting strategy on the first column of the reduced matrix, we take $Q = I$ in the above.)

2.6 Symmetric Case : Direct Methods

If A is symmetric, then we can only apply congruences to A if we want to preserve symmetry. In particular, whenever we interchange two rows, we must also interchange the corresponding columns. Thus only a diagonal element can be brought into the leading diagonal position. The symmetric form of the Gaussian elimination decomposition $L\,U$ gives the decomposition $L\,D\,L^t$, where L is unit lower triangular, D is diagonal, and L^t is the transpose of L . Since we may only perform congruences on A , the error matrix is $E = E(L, D, A, b, N, N^t)$ where N is a permutation matrix such that $L\,D\,L^t = N\,A\,N^t$.

2.7 Symmetric Positive Definite Case

(a) Cholesky's Method

The Cholesky decomposition (Wilkinson (1965), pp. 229-232) is the most well-known decomposition for a positive definite matrix A

(i.e. $x^t A x > 0$ for $x \neq 0$). Here $A = \tilde{L} \tilde{L}^t$, where $\tilde{L}$ is lower triangular. No interchanges are required for stability.

(b) $L D L^t$ Decomposition

If A is positive definite, then its $L D L^t$ decomposition (the symmetric form of Gaussian elimination) is stable in the absence of underflow and overflow. The elements of L can be arbitrarily large, but if they do not overflow then in fact the error matrix E is as small as the error matrix for the Cholesky decomposition of A.
(Note: $\tilde{L} = L D^{1/2}$).

(c) Method of Congruent Transformations

This method (Westlake, p. 21; De Meersman and Schotsmans, (1964)) uses decomposition (b) with interchanges. Here the largest diagonal element is brought into the leading diagonal position at each step. If A is positive definite, then the elements of L are bounded by 1 and the method is stable.

(d) Summary

If A is positive definite, then the above three methods are stable. If A is $n \times n$, then each method requires $\sim \frac{1}{2} n^2$ storage positions, $\sim \frac{1}{6} n^3$ multiplications, and $\sim \frac{1}{6} n^3$ additions to solve $A x = b$.

2.8   Symmetric Case : <u>Failure of Cholesky</u> and $L D L^t$ <u>Methods in Exact Arithmetic</u>

If A is symmetric but indefinite, then the $L D L^t$ decomposition, the method of congruent transformations, and the Cholesky decomposition fail in exact arithmetic for a matrix as simple as

$A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$, that is, there exists no permutation matrix N such that

$N A N^t$ has an $L D L^t$ or an $\tilde{L} \tilde{L}^t$ decomposition.

Note that $\tilde{L} \tilde{L}^t$ is always positive semi-definite. Thus Cholesky decomposition will fail in exact arithmetic whenever A is indefinite and $\det A \neq 0$, i.e. there exists no permutation N such that $N A N^t = \tilde{L} \tilde{L}^t$ where $\tilde{L}$ is lower triangular.

The $L D L^t$ decomposition and the method of congruent transformations will fail in exact arithmetic whenever all the diagonal elements in a reduced matrix are zero, i.e. there exists no permutation N such that $N A N^t$ has an $L D L^t$ decomposition.

2.9 Symmetric Case : Failure of $L D L^t$ in Finite Precision Arithmetic

The $L D L^t$ decomposition in finite precision can be unstable if the diagonal elements are too small. The $L D L^t$ decomposition on the

matrix $A = \begin{bmatrix} \epsilon & 1 \\ 1 & \eta \end{bmatrix}$ will produce the same incorrect solution on the

computer as we saw in §2.3 for Gaussian elimination without interchanges. However, here there exists no permutation matrix N such that $N A N^t$ has a stable $L D L^t$ decomposition. Thus the $L D L^t$ decomposition fails for symmetric indefinite matrices.

2.10 Symmetric Case : Present Situation

If we ignore the symmetry of A and apply elimination with complete or partial pivoting to A, then in general A will no longer be symmetric after the first step of the elimination. We then need $\sim n^2$ storage positions in the computer and we must perform $\sim \frac{1}{3} n^3$ multiplications and $\sim \frac{1}{3} n^3$ additions. But we also have stability for the

L U decomposition. This procedure is presently recommended for the solution of symmetric indefinite systems of linear equations (Fox; p. 80, 185), and thus the symmetry of A is of no advantage.

2.11   Symmetric Case : Our Problem

If A is symmetric but indefinite, we would like to find an algorithm which gives a stable decomposition when applied to $N A N^t$ where N is a suitable permutation matrix, but which also takes advantage of the symmetry in order to require only $\sim \frac{1}{2} n^2$ storage positions in the computer and to require only $\sim \frac{1}{6} n^3$ multiplications and $\sim \frac{1}{6} n^3$ additions.

Our algorithm will fulfill all the above requirements with the exception that we will need between $\sim \frac{1}{4} n^3$ and $\sim \frac{1}{3} n^3$ additions.

Chapter 3 : <u>Historical Survey</u>

3.1   <u>Introduction</u>

Various methods have been proposed for symmetric indefinite matrices. Most of these methods have been unstable, while the stable methods have required operation counts of at least $n^3/3$, where an operation is defined to be a multiplication followed by an addition. Let us look at some of these methods.

3.2   <u>Direct Methods</u>

Usually direct methods for symmetric indefinite systems are based on the symmetric form of Gaussian elimination, $L D L^t$, which is unstable in the absence of pivoting.

The $L D L^t$ method and its variant, the method of congruent transformations (in which we use the largest diagonal element as the pivot at each step (§2.7)) require $\sim \frac{1}{2} n^2$ storage locations and $\sim \frac{1}{6} n^3$ operations. But both methods are unstable (§2.7-2.9). These methods are of value only if it is known in advance that no element of D will vanish or be small.

The Crout factorization (Hildebrand, pp. 429-435; Householder, pp. 82-83) is also a modification of $L D L^t$, symmetric Gaussian elimination, and thus requires $\sim \frac{1}{2} n^2$ storage locations and $\sim \frac{1}{6} n^3$ operations, but it is also unstable.

These variants of $L D L^t$ are unstable. Let us consider some direct methods that are not based on $L D L^t$.

The escalator method (Householder, pp. 78-79) uses the known solution of a subsystem as a step in solving the complete system. The problem lies in finding the solution of the subsystem.

Some headway in this problem was made by Parlett and Reid (1969). They reduce a symmetric matrix to tridiagonal form by stabilized elementary congruences and solve the tridiagonal system by Gaussian elimination with partial pivoting. They require $\sim \frac{1}{2} n^2$ storage locations and $\sim \frac{1}{3} n^3$ operations, and the method is stable.

In October, 1965, W. Kahan (in correspondence with R. De Meersmans and L. Schotsmans) proposed a method for solving symmetric systems based on Lagrange's theorem on the reduction of quadratic forms to diagonal forms (§4.3 - 4.4). Kahan proposed the generalization of the idea of a pivot to include 2×2 submatrices (§4.5).

Then Schotsmans (1965) prepared an algorithm in which one searches all the principal 2×2 submatrices for the one with largest determinant. This algorithm requires $\sim \frac{1}{2} n^2$ storage, but between $\sim \frac{1}{3} n^3$ and $\sim \frac{1}{2} n^3$ operations (§5.4).

3.3   Indirect Methods

The Seidel iterative method (Householder, pp. 48-51, 81) and the method of relaxation (Householder, pp. 48-51, 81) require $\sim n^2$ operations for each cycle. The number of cycles required depends on the matrix, the starting values, and the needed accuracy. Usually the number of cycles exceeds n, so at least $\sim n^3$ operations are required.

The method of steepest descent (Householder, pp. 47-51, 82) requires $\sim 2 n^2$ operations at each step. Again, usually at least n steps are required. So the number of operations is at least $\sim 2 n^3$.

The congruent gradient method, also called the Stiefel-Hestenes method, is a finite iterative method designed for positive definite matrices, but it can be used for symmetric matrices (Fox, pp. 208-213; Householder, pp. 73-78, 82). It requires $\sim 2 n^2$ operations at each of n steps. Once again $\sim 2 n^3$ operations are required. See Reid (1967) for useful observations on this method.

Chapter 4 : <u>Diagonal Pivoting</u>

4.1 <u>Preserving Symmetry</u>

In order to have a direct method for symmetric matrices which will preserve symmetry, we can perform only congruences on the matrix A , i.e. if we premultiply A by a non-singular matrix X , then we must also postmultiply by X^t .

4.2 The $L D L^t$ <u>Decomposition</u>

Let us consider the $L D L^t$ method in greater detail. We convert A to diagonal form by congruences. Let us consider the first step of the decomposition:

Let $A = \begin{bmatrix} a & c^t \\ \hline c & B \end{bmatrix}$. If $a \neq 0$, then $A = L \begin{bmatrix} a & 0 \\ \hline 0 & B-CC^t/a \end{bmatrix} L^t$,

where $L = \begin{bmatrix} 1 & 0 \\ \hline C/a & I_{n-1} \end{bmatrix}$ and I_{n-1} is the identity matrix of order $n - 1$.

The variant of $L D L^t$ called the method of congruent transformations (Westlake, p. 21; De Meersmans and Schotsmans) uses the largest diagonal element as pivot. This is equivalent to the $L D L^t$ decomposition of $N A N^t$, where N is a permutation matrix.

As we saw in §2.8 both methods are unstable for symmetric (indefinite) systems.

This instability results from our being unable to bring an off-diagonal element into the pivotal position. Since we are using only congruences, we can bring only a diagonal element into the pivotal

position. When some off-diagonal element A_{ji} , $j > 1$, is very large, we can bring A_{ji} into the $(2,1)$ position by congruences, but never into the $(1,1)$ position. Thus we cannot take advantage of this valuable information.

4.3   Orthogonal Reduction to Diagonal Form

Any real quadratic form $x^t A x$ of rank r can be reduced by an orthogonal transformation to a diagonal form

$$\lambda_1 x_1^2 + .. + \lambda_r x_r^2 ,$$

where $\lambda_1, \ldots, \lambda_r$ are the non-zero eigenvalues of A (Mirsky, pp. 362-363).

If A is an $n \times n$ symmetric matrix with $\det A \neq 0$, then the above means that

$$A = 0 \, \Lambda \, 0^t ,$$

where $\Lambda = \text{diag } \lambda_1, \ldots, \lambda_n$, the λ_i are the eigenvalues of A , and 0 is an orthogonal matrix whose i^{th} column is an eigenvector corresponding to λ_i .

However, this $0 \, \Lambda \, 0^t$ decomposition involves more work than Gaussian elimination and requires the use of irrational operations.

For a finite-precision algorithm we would prefer a reduction involving only rational operations.

4.4   Lagrange's Method of Reduction

In 1759 Lagrange devised a method for reducing a real quadratic form of rank r by a real non-singular linear transformation to a

diagonal form

$$\alpha_1 x_1^2 + \ldots + \alpha_r x_r^2 ,$$

where $\alpha_1, \ldots, \alpha_r$ are all non-zero (Mirsky, pp. 368-374), and the number of positive (and negative) squares is invariant.

This method corresponds to the $L\,D\,L^t$ decomposition of a symmetric matrix A when the $L\,D\,L^t$ decomposition exists.

Suppose $A_{11} = \ldots = A_{nn} = 0$, while $\det A \neq 0$. Then the $L\,D\,L^t$ decomposition for A does not exist. In this case, some $A_{rs} \neq 0$ for $r \neq s$ since $\det A \neq 0$.

Let us assume $A_{11} = 0 = A_{22}$ but $A_{12} \neq 0$, where $A = A^t$. $\phi(x_1, \ldots, x_n) = x^t A x$ is a quadratic form in $x_1, \ldots, x_n$ where $x^t = [x_1, \ldots, x_n]$.

In this case Lagrange proposed the following transformation:

(4.4.1) $\quad x_1 = y_1 + y_2$, $x_2 = y_1 - y_2$, $x_3 = y_3, \ldots, x_n = y_n$.

This maps $2\,A_{12}\,x_1\,x_2$ into $2\,A_{12}\,(y_1^2 - y_2^2)$. Thus ϕ is transformed into a quadratic form ψ in $y_1, \ldots, y_n$ where the coefficients of y_1^2 and y_2^2 are non-zero. Then we can proceed with the decomposition (Mirsky, p. 371-2; Gantmacher, p. 199).

Let us consider the above transformation in matrix form. Then $T\,y = x$, where

(4.4.2) $\quad T = \left[\begin{array}{cc|c} 1 & 1 & 0 \\ 1 & -1 & \\ \hline 0 & & I_{n-2} \end{array}\right]$ and I_{n-2} is the identity matrix of order

$n - 2$.

Thus $x^t A x = y^t (T^t A T) y$, and $T^t A T$ is a symmetric matrix with $(T^t A T)_{11}, (T^t A T)_{22} \neq 0$.

Hence we have avoided the problem of zeros on the diagonal of A by use of the 2×2 matrix $\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$.

This procedure is also applicable to complex quadratic forms.

4.5 Kahan's Proposal

In 1965 W. Kahan (in correspondence with R. De Meersmans and L. Schotsmans) proposed that Lagrange's method could be made the basis of a stable method which preserved symmetry.

Kahan adapted Lagrange's method to finite precision by observing that the use of $\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$ in (4.4.2) corresponds to the use of a 2×2 submatrix as a pivot in a decomposition by linear transformations and that a 2×2 could be chosen if the diagonal elements were zero or very small (§§2.8-2.9).

Suppose we used a 2×2 submatrix P as a pivot. Let us look at such a decomposition. Let $A = \left[\begin{array}{c|c} P & C^t \\ \hline C & B \end{array}\right]$, where $A = A^t$, det $A \neq 0$, A is $n \times n$, P is 2×2 , C is $(n-2) \times 2$, and B is $(n-2) \times (n-2)$. Then $A = L_1 \left[\begin{array}{c|c} P & 0 \\ \hline 0 & B - CP^{-1} C^t \end{array}\right] L_1^t$, where $L_1 = \left[\begin{array}{c|c} I_2 & \\ \hline CP^{-1} & I_{n-2} \end{array}\right]$ and I_k is the identity matrix of order k .

How do we choose whether to use a 1×1 or a 2×2 pivot? Can both 1×1 and 2×2 pivots be bad?

4.6   Pivotal Strategy

Kahan considered two pivotal strategies. In the first, one searches the entire matrix for $m_c = \max_{i,j,k} \{A_{ii}^2, |A_{jj} A_{kk} - A_{jk}^2|\}$.

If $m_c = A_{ii}^2$, then interchange rows and columns 1 and i and use A_{ii} as a 1×1 pivot. If $m_c = |A_{jj} A_{kk} - A_{kj}^2|$, then interchange rows and columns 1 with j and 2 with k, and use $\begin{bmatrix} A_{jj} & A_{jk} \\ A_{jk} & A_{kk} \end{bmatrix}$ as a 2×2 pivot.

Since we search the entire matrix, this is called a complete pivoting strategy, in analogy with complete pivoting for Gaussian Elimination. However, the searching here requires between $\sim \frac{1}{6} n^3$ and $\sim \frac{1}{3} n^3$ multiplications to find m_c for all steps (depending on the number of 1×1 and 2×2 pivots used). The decomposition itself requires $\sim \frac{1}{6} n^3$ multiplications. Thus this strategy would require between $\sim \frac{1}{3} n^3$ and $\sim \frac{1}{2} n^3$ multiplications to solve $A x = b$ (§5.4), which is more than for Gaussian elimination. Hence Kahan rejected this strategy.

Kahan considered a second pivotal strategy in which we scan only the first column and the main diagonal; this is called a partial pivoting strategy, in analogy with partial pivoting for Gaussian elimination. The searching here only requires between $\sim \frac{1}{4} n^2$ and $\sim \frac{1}{2} n^2$ multiplications.

We take $m_p = \max_{i,j} \{A_{ii}^2, |A_{11} A_{jj} - A_{1j}^2|\}$. However, this partial pivoting strategy is unstable.

$$\text{Let } A = \begin{bmatrix} \frac{1}{2}\epsilon & \epsilon & \epsilon \\ \epsilon & \frac{1}{2}\epsilon & 1 \\ \epsilon & 1 & \frac{1}{2}\epsilon \end{bmatrix} \text{ , where } 0 < \epsilon \ll 1 \text{ .}$$

Then $m_p = |A_{11} A_{22} - A_{21}^2| = \frac{3}{4} \epsilon^2$. Thus $\begin{bmatrix} \frac{1}{2}\epsilon & \epsilon \\ \epsilon & \frac{1}{2}\epsilon \end{bmatrix}$ would be used as

a 2×2 pivot, and the reduced matrix $A^{(1)}$ is $[-\frac{2}{3}\frac{1}{\epsilon} + \frac{8}{3} - \frac{2}{3}\epsilon]$.

If ϵ is small enough, then in finite precision arithmetic the

operation $-\frac{2}{3}\frac{1}{\epsilon} + \frac{8}{3} - \frac{2}{3}\epsilon$ yields $-\frac{2}{3}\frac{1}{\epsilon} + \frac{8}{3}$. This can cause highly

inaccurate solutions, as in §2.3.

So this partial pivoting strategy is unstable. For these reasons

Kahan rejected this method for use on symmetric systems.

4.7   Parlett's Observation

In 1967 B. Parlett observed that the examples for which the partial

pivoting strategy was unstable were also unequilibrated. A symmetric

matrix A is equilibrated if $\max_j |A_{ij}| = 1$ for each row index i

(§7.1). Parlett conjectured that the partial diagonal pivoting strategy

would be stable when applied to equilibrated matrices.

Chapter 5 : The Decomposition for Diagonal Pivoting

5.1   Definitions

Let A be an $n \times n$ symmetric non-singular matrix. We want to

reduce A to the "diagonal" form $M D M^t$ by congruences, where D

is a block diagonal matrix, each block being of order 1 or 2, and M

is unit lower triangular with $M_{i+1,i} = 0$ if $D_{i+1,i} \neq 0$.

Let $\mu_0 = \max_{i,j} |A_{ij}|$, $\mu_1 = \max_{i} |A_{ii}|$, and $\nu = |A_{11} A_{22} - A_{21}^2|$.

5.2   The Decomposition

Let $A = \left[\begin{array}{c|c} P & C^t \\ \hline C & B \end{array} \right]$, where C is $j \times (n-j)$, B is $(n-j) \times (n-j)$,

and P is $j \times j$, where $j = 1$ or 2.

If P^{-1} exists, then $A = L_1 \left[\begin{array}{c|c} P & 0 \\ \hline 0 & B - C P^{-1} C^t \end{array} \right] L_1^t$, where

$L_1 = \left[\begin{array}{c|c} I_j & 0 \\ \hline C P^{-1} & I_{n-j} \end{array} \right]$, and I_j, I_{n-j} are the identity matrices of order

j and n-j, respectively. Any element of $C P^{-1}$ will be called a

multiplier.

5.3   The 1×1 Pivot

Suppose P is of order 1. (We shall not make a distinction

between a matrix P of order 1 and its element, which we shall also

call P .)

Let us assume that we have already interchanged rows and columns so that $|P| = \mu_1$, i.e. P is the maximum diagonal element.

If P^{-1} exists (i.e. $\mu_1 \neq 0$), then let $A^{(n-1)} = B - C P^{-1} C^t$.

Then $(C P^{-1})_i = A_{i+1}/P$ and $A_{ij}^{(n-1)} = A_{i+1,j+1} - (C P^{-1})_i A_{j+1,1}$.

Since $|P| = \mu_1$ and $\mu_0 = \max_{i,j} |A_{ij}|$, we have the following:

Lemma 1: If P is of order 1 and $|P| = \mu_1 \neq 0$, then

(i) $\max_i |(C P^{-1})_i| \leq \mu_0/\mu_1$,

(ii) $\max_{i,j} |A_{ij}^{(n-1)}| \leq (1 + \mu_0/\mu_1)\mu_0$.

Thus a 1×1 pivot P is useful iff $|P| = \mu_1$ is large relative to μ_0, i.e. if μ_1/μ_0 is bounded away from zero.

5.4   The 2×2 Pivot

Suppose P is of order 2 and P^{-1} exists (i.e. $\nu \neq 0$).
Here the $(k-1)^{st}$ row of $C P^{-1}$ is:

$$(A_{k1}, A_{k2}) P^{-1} = \frac{1}{A_{11} A_{22} - A_{21}^2} (A_{k2} A_{22} - A_{k2} A_{21}, A_{k2} A_{11} - A_{k1} A_{21}).$$

let $A^{(n-2)} = B - C P^{-1} C^t$. Then $A_{ij}^{(n-2)} = A_{i+2,j+2} - (C P^{-1})_{i1} C_{1j} - (C P^{-1})_{i2} C_{2j}$.

Since $\nu = |A_{11} A_{22} - A_{21}^2|$, $|A_{k1}|$ and $|A_{k2}| \leq \mu_0$, and $|A_{11}|, |A_{22}| \leq \mu_1$, we have the following:

Lemma 2: If P is of order 2 and $|\det P| = \nu \neq 0$, then

(i) $\max\limits_{i,j} |(C\,P^{-1})_{ij}| \leq \mu_0\,(\mu_0 + \mu_1)/\nu$,

(ii) $\max\limits_{i,j} |A_{ij}^{(n-2)}| \leq [1 + 2\mu_0\,(\mu_0 + \mu_1)/\nu]\,\mu_0$.

Thus a 2×2 pivot is useful iff we can bound ν away from zero. In particular from §5.3, we need to have ν bounded away from zero whenever μ_1/μ_0 is near zero.

(Note that the use of the standard norm bound would give too crude an analysis for Lemmas 1 and 2.)

5.5   Bounding ν

We can easily bound ν from above, since $\nu = |A_{11}\,A_{22} - A_{21}^2| \leq |A_{21}|^2 + |A_{11}|\,|A_{22}| \leq \mu_0^2 + \mu_1^2$. Thus we have:

Lemma 3: $|\det P| = \nu \leq \mu_0^2 + \mu_1^2$.

This upper bound is sharp for

$$
A = \begin{bmatrix}
\mu_1 & \mu_0 & & & 0 \\
\mu_0 & -\mu_1 & & & \\
& \mu_0 & -\mu_1 & & \\
& & & \mu_0 & \\
& & & -\mu_1 & \\
0 & & & & \mu_0
\end{bmatrix}
$$

But, as we saw in §5.4, we need a lower bound on ν which bounds

ν away from zero when μ_1/μ_0 is small. Clearly, such a lower bound does not exist without interchanges.

$$\text{Consider} \quad A = \begin{bmatrix} \epsilon & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & \epsilon \end{bmatrix} \quad \text{with} \quad P = \begin{bmatrix} \epsilon & 0 \\ 0 & 0 \end{bmatrix} .$$

We shall exhibit three different pivotal strategies in §6.1, §6.4, and §8.1, which provide us with the necessary interchanges so that we have a 2×2 pivot P with:

$$|\det P| = \nu \geq \mu_0^2 - \mu_1^2 . \quad (\S 6.2, \S 6.5, \S 8.2).$$

Assuming this lower bound, we have:

<u>Lemma 4</u>: If $|\det P| = \nu \geq \mu_0^2 - \mu_1^2 > 0$, then

(i) $\displaystyle\max_{i,k} |(C\,P^{-1})_{ik}| \leq \mu_0/(\mu_0 - \mu_1)$ for $k=1,2,$

(ii) $\displaystyle\max_{i,j} |A_{ij}^{(n-2)}| \leq \mu_0 [1 + 2\mu_0/(\mu_0 - \mu_1)] .$

The lower bound on $\nu > 0$ is sharp for

$$A = \begin{bmatrix} \mu_1 & \mu_0 & & & 0 \\ \mu_0 & & & & \\ & & & & \mu_0 \\ 0 & & & \mu_0 & \mu_1 \end{bmatrix}$$

Thus we have a good bound on element growth in the reduced matrix, since if μ_1/μ_0 is small then $1 - \mu_1/\mu_0 \doteqdot 1$. We shall see in Chapter 10–12 that stability follows from this.

5.6   The Reduced Matrices

Define $A^{(n)} = A$, $\mu_0^{(n)} = \mu_0$, $\mu_1^{(n)} = \mu_1$, $\nu^{(n)} = \nu$.

Let $A^{(k)}$ be the reduced matrix of order k . Let

$$\mu_0^{(k)} = \max_{i,j} |A_{ij}^{(k)}| \; , \; \mu_1^{(k)} = \max_j |A_{jj}^{(k)}| \; , \text{ and } \; \nu^{(k)} = |A_{11}^{(k)} A_{22}^{(k)} - A_{21}^{(k)2}| \; .$$

All considerations in §§5.2-5.5 hold for $A^{(k)}$.

5.7   Criterion for Choosing a 1×1 or 2×2 Pivot

We must find a proper criterion for deciding whether we shall use a 1×1 or 2×2 pivot.

In Chapter 10 we shall show that the elements of the error matrix are bounded in proportion to the elements in the reduced matrices. For stability we must ensure that the elements in the reduced matrices do not become too large.

If we made our criterion to be the minimization of the number of multiplications (additions), then we would want a 1×1 (2×2) pivot at each step. But this would be unstable.

Instead, let us aim to minimize the element growth that can take place in the transformation from one reduced matrix to the next. For further remarks see §12.6.

Let $F_j^{(k)}$ be the growth factor permitted by choosing a $j×j$ pivot for $A^{(k)}$, where $j=1$ or 2.

If the hypothesis of Lemma 4 holds (i.e. $\nu^{(k)} \geq \mu_0^{(k)2} - \mu_1^{(k)2}$ for all $A^{(k)}$), then by Lemmas 1 and 4:

$$F_1^{(k)} = 1 + \mu_0^{(k)}/\mu_1^{(k)} \quad , \quad F_2^{(k)} = 1 + 2/(1 - \mu_1^{(k)}/\mu_0^{(k)}) \ .$$

$F_1^{(k)}$ has a good bound if $\mu_1^{(k)}/\mu_0^{(k)}$ is not too small; while $F_2^{(k)}$ has a good bound if $\mu_1^{(k)}/\mu_0^{(k)}$ is not too large. Thus we are led to the following:

Definition: For $0 < \alpha < 1$, let S_α be the following strategy: for each reduced matrix $A^{(k)}$, choose a 1×1 pivot iff $\mu_1^{(k)}/\mu_0^{(k)} \geq \alpha$ (and a 2×2 pivot otherwise).

With S_α we have $F_1^{(k)} \leq 1 + 1/\alpha$ and $F_2^{(k)} \leq 1 + 2/(1 - \alpha)$ for all $A^{(k)}$.

But at any stage the choice of a 2×2 pivot carries us further towards the complete reduction than does the choice of a 1×1 pivot. Since the growth factors from reduced matrix to reduced matrix are multiplicative, it is natural to compare the square of the growth factor $F_1^{(k)}$ permitted by choosing a 1×1 pivot for $A^{(k)}$ with the growth factor $F_2^{(k)}$ for a 2×2 pivot.

Thus the problem is to find $\min_{0<\alpha<1} \max \{(1 + 1/\alpha)^2, 1 + 2/(1 - \alpha)\}$.

Theorem: $\min_{0<\alpha<1} \max \{(1 + 1/\alpha)^2, 1 + 2/(1 - \alpha)\} = (9 + \sqrt{17})/8$

and is achieved by $\alpha = \alpha_0 = (1 + \sqrt{17})/8$.

Proof: The equation $(1 + 1/\alpha)^2 = 1 + 2/(1 - \alpha)$ reduces to a quadratic with roots $(1 \pm \sqrt{17})/8$. Since the left side of the equation

is monotone decreasing, the right side is monotone increasing, and $\alpha > 0$, the minimum is given by $\alpha = \alpha_0 = (1 + \sqrt{17})/8$. $q.e.d.$

We immediately obtain the following bounds on multipliers and on elements in the reduced matrices under strategy S_α . Let m be any multiplier.

<u>Corollary 2</u>: Under strategy S_α , if for all $A^{(k)}$, $\nu^{(k)} \geq \mu_0^{(k)2} - \mu_1^{(k)2}$, then:

$$|m| \leq \begin{cases} 1/\alpha & \text{for a } 1\times 1 \text{ pivot} \\ 1/(1 - \alpha) & \text{for a } 2\times 2 \text{ pivot} \end{cases} .$$

For $\alpha = \alpha_0 = (1 + \sqrt{17})/8$,

$$|m| \leq \begin{cases} (\sqrt{17} - 1)/2 < 1.562 & \text{for a } 1\times 1 \text{ pivot} \\ (\sqrt{17} + 7)/4 < 2.781 & \text{for a } 2\times 2 \text{ pivot} \end{cases}$$

<u>Corollary 3</u>: Under strategy S_α with $\alpha = \alpha_0$, if $\nu^{(k)} \geq \mu_0^{(k)2} - \mu_1^{(k)2}$ for all $A^{(k)}$, then for $1 \leq i \leq n$:

$$\mu_0^{(i)} \leq \mu_0[(9 + \sqrt{17})/2]^{(n-i)/2} < \mu_0 (2.57)^{n-i} .$$

In Chapters 11-12 we will give a much better bound on the elements in the reduced matrices.

The strategy S_α allows us to proceed in the following order:

(1) calculate $\mu_0^{(k)}$ and $\mu_1^{(k)}$;

(2) if $\mu_1^{(k)} \geq \alpha_0 \mu_0^{(k)}$ then we use a 1×1 pivot;

(3) otherwise we find a 2×2 by some strategy.

Thus we search for a 2×2 for $A^{(k)}$ iff $\mu_1^{(k)} < \alpha_0 \mu_0^{(k)}$.

Chapter 6 : <u>The Complete and Partial Pivoting Strategies</u>

6.1   <u>Complete Pivoting</u>

As we saw in Chapter 4, the partial pivoting strategy can be
unstable when used on unequilibrated matrices. The trouble lies in
the fact that we do not have a lower bound on the 2×2 principal minors
which bounds them away from zero when the diagonal elements are small
(§5.5).

Let us therefore consider a complete pivoting strategy ("complete"
in the sense that we search over all the principal 2×2 minors, cf. §4.6).

By interchanging rows and the corresponding columns it is possible
to bring any diagonal element into the (1,1) position or any principal
2×2 submatrix into the leading 2×2 position.

Let $\nu_c = \max_{i,j} |A_{11} A_{jj} - A_{ij}^2|$.

The complete strategy involves:

(1)   finding $\mu_1 = \max_1 |A_{ii}|$, $\mu_0 = \max_{i,j} |A_{ij}|$;

(2)   choosing a 1×1 or 2×2 pivot according to S_α (§5.7);

(3)   for a 1×1, interchanging so that $|P| = \mu_1$ (§5.3);

(4)   for a 2×2, finding ν_c , and interchanging so that $|\det P| = \nu_c$
      (§5.4).

This would be repeated for each reduced matrix.

6.2   Bounding ν_c

However, the result of all this work is that we do obtain a lower
bound for ν_c in terms of μ_0 and μ_1 .

Theorem 1: $\mu_0^2 - \mu_1^2 \leq \nu_c \leq \mu_0^2 + \mu_1^2$.

Proof: The upper bound follows from Lemma 3 of §5.5. Let
$\mu_0 = |A_{rs}|$. Then $\nu_c = \max_{i,j} |A_{ii} A_{jj} - A_{ij}^2| \geq |A_{rr} A_{ss} - A_{rs}^2| =$

$\mu_0^2 - A_{rr} A_{ss} \geq \mu_0^2 - \mu_1^2$, since $\mu_0^2 = A_{rs}^2$ and

$\mu_1 = \max_i |A_{ii}|$.   q.e.d.

Thus Lemma 4 in §5.5 holds for ν_c when $\mu_1 < \mu_0$. According to
S_α , we choose a 1×1 pivot for the reduced matrix $A^{(k)}$ if

$\mu_1^{(k)} \geq \alpha_0 \mu_0^{(k)}$, where $\alpha_0 = (1 + \sqrt{17})/8$.

In Chapters 10-12 we shall see that stability of this complete
pivoting strategy follows.

6.3   Operation Count for Complete Pivoting

Unfortunately, the operation count here is much larger than we
desire.

The calculation of $\nu_c^{(k)}$ requires $k(k-1)$ multiplications and
additions. Let p be the number of 1×1 pivots used (so $q = (n - p)/2$
pivots of order 2 are used). Let $\sum^{(j)}$ denote summation over those
indices k , $1 \leq k \leq n$, for which $A^{(k)}$ uses a pivot of order j .

The searching for all the $v_c^{(k)}$ requires:

(6.3.1) $\sum^{(1)} k(k-1) + \sum^{(2)} k(k-1)$ multiplications and additions.

Now (6.3.1) $\leq \sum_{r=1}^{n} k(k-1) = \frac{1}{3} n(n+1)(n-4) \sim \frac{1}{3} n^3$ with

equality iff $p = n$, while (6.3.1) $\geq \sum_{j=1}^{n/2} 2j(2j-1) = \frac{1}{12} n(n+2)(2n-1)$

$\sim \frac{1}{6} n^3$ with equality iff $p = 0$ (i.e. $q = n/2$).

From Chapter 9 we see that the rest of the work for solving $A x = b$ would require $\sim \frac{1}{6} n^3$ multiplications, and between $\sim \frac{1}{4} n^3$ and

$\sim \frac{1}{3} n^3$ additions.

Thus the complete diagonal pivoting strategy requires between

$\sim \frac{1}{3} n^3$ and $\sim \frac{1}{2} n^3$ multiplications, and between $\sim \frac{5}{12} n^3$ and $\sim \frac{2}{3} n^3$

additions. (To be exact, $\sim \frac{1}{3} n^3 + \frac{1}{6} p^3$ multiplications and

$\sim \frac{5}{12} n^3 + \frac{1}{4} p^3$ additions are required, where p is the number of 1×1

pivots used.)

Examples: If A is $n \times n$ and positive definite, then $p = n$.

If $A =$ $\begin{bmatrix} 0 & 1 & & & 0 \\ 1 & & & & \\ & & & & 1 \\ 0 & & & 1 & 0 \end{bmatrix}$ is $n \times n$ with n even, then $p = 0$.

6.4 Partial Pivoting for Equilibrated Matrices

As we saw in §§6.2-6.3 the complete pivoting strategy is stable, but it involves more work that we are willing to perform, while in Chapter 3 we saw that a partial pivoting strategy is unstable for unequilibrated matrices.

We shall now show that a partial pivoting strategy is stable when applied to equilibrated matrices (with equilibrated reduced matrices).

Let A be equilibrated, i.e. let $\max_j |A_{ij}| = \mu_0$ for every i, where $\mu_0 > 0$ (usually we normalize by taking $\mu_0 = 1$).

Let $\nu_p = \max_i |A_{11} A_{ii} - A_{ii}^2|$.

The partial strategy involves:

(1) equilibrating A (thus we know μ_0);

(2) finding μ_1 and choosing a 1×1 or 2×2 according to S_α (§5.7);

(3) for a 1×1, interchanging so that $|P| = \mu_1$ (§5.3);

(4) for a 2×2, finding ν_p , and interchanging so that $|\det P| = \nu_p$ (§5.4).

6.5 Bounding ν_p

Now let us find a lower bound for ν_p .

__Theorem 2__: If A is equilibrated ($\max_j |A_{ij}| = \mu_0$ for every i), then $\mu_0^2 - \mu_1^2 \leq \nu_p \leq \mu_0^2 + \mu_1^2$.

Proof: The upper bound follows from Lemma 3 of §5.5 By

equilibration, either (i) $|A_{11}| = \mu_0$, or (ii) there exists integer

$k \geq 2$ with $|A_{k1}| = \mu_0$. If (i), then $\mu_0 = \mu_1$ and, trivially,

$v_p \geq \mu_0^2 - \mu_1^2 = 0$. If (ii), then $v_p \geq |A_{11} A_{kk} - A_{k1}^2| =$

$\mu_0^2 - A_{11} A_{kk} \geq \mu_0^2 - \mu_1^2$. $q.e.d.$

Thus, Lemma 4 in §5.5 holds for v_p if $\mu_1 < \mu_0$. According to

S_α, we choose a 1×1 pivot for the equilibrated reduced matrix $A^{(k)}$

of order k iff $\mu_1^{(k)} \geq \alpha_0 \mu_0^{(k)}$, where $\alpha_0 = (1 + \sqrt{17})/8$.

Also, stability of the partial pivoting strategy for equilibrated

matrices follows from Chapters 10-12.

For $A = A^{(n)}$, only $2(n - 1)$ multiplications and additions are

required to calculate $v_p^{(n)}$. Thus the calculation of $v_p^{(k)}$ for all

k requires between $\frac{1}{2} n^2$ and $n(n - 1)$ multiplications, compared

with between $\sim \frac{1}{6} n^3$ and $\sim \frac{1}{3} n^3$ for all the $v_c^{(k)}$ in (6.3.1).

6.6   Criticism of the Partial Pivoting Strategy

The drawback to this method and the criterion which we have found

for the pivoting strategy is that the matrix must be equilibrated at

the beginning and then each reduced matrix should be equilibrated.

But an algorithm for equilibrating symmetric matrices has never been

exhibited, i.e., for an arbitrary matrix A, we seek diagonal matrices

D_1, D_2 such that $D_1 A D_2$ is equilibrated, and if A is symmetric,

we need $D_1 = D_2$ in order to preserve symmetry.

We resolve the above predicament by two fundamentally different approaches. In Chapter 7 we exhibit an algorithm which can equilibrate any symmetric matrix in a very simple way. In §§6.1-6.3 we showed that complete diagonal pivoting avoids the problem of equilibration and is stable, although the number of multiplications and additions required is more than we desire. But in Chapter 8 we shall exhibit a new version of diagonal pivoting which is applicable to unequilibrated matrices; we call this unequilibrated diagonal pivoting. This method will show that equilibration (in Wilkinson's sense) is unnecessary for this strategy and this is the algorithm that we recommend.

Chapter 7 : <u>Equilibration of Symmetric Matrices</u>

7.1   <u>Introduction</u>

Wilkinson (1961) recommends that a matrix be equilibrated before applying any algorithm for solving a system of linear equations. A matrix A is said to be equilibrated if all its rows and columns have the same length in some norm. Wilkinson's rounding error analysis for Gaussian elimination (Wilkinson, 1961) gives the most effective results when the matrix is equi'ibrated, since a small perturbation of one row (or column) is then of the same magnitude as that of any other row (or column).

7.2   <u>Equilibration of General Matrices</u>

In finite precision we modify the definition of equilibration. (In this chapter we shall confine ourselves to the norm

$$\|x\|_\infty = \max_i |x_i| \quad .)$$

A matrix A is row equilibrated if, for each row index i ,

$\beta^{-1} \mu_0 \leq \max_{1 \leq j \leq n} |A_{ij}| \leq \mu_0$, where β is the number base of the floating

point system. A matrix A is column equilibrated if, for each column

index j , $\beta^{-1} \mu_0 \leq \max_{1 \leq j \leq n} |A_{ij}| \leq \mu_0$.

A matrix A is equilibrated if it is both row and column equili-brated.

Usually we normalize by choosing $\mu_0 = 1$. We shall assume $\mu_0 = 1$ in this chapter.

The use of β permits a matrix to be equilibrated by changes of exponent only.

In order to row (column)-equilibrate A we seek a diagonal matrix D_1 (D_2) such that $D_1 A$ $(A D_2)$ is row (column)-equilibrated. To equilibrate A we seek diagonal matrices D_1, D_2 such that $D_1 A D_2$ is equilibrated. However, there is no unique equilibrated form of a matrix for this norm (Forsythe and Moler, p. 45). The various equilibrated forms may differ greatly in their desirability for use in Gaussian elimination, since the various equilibrations cause different choices of the pivots (some are good choices, others are bad).

For the one-norm $(\|x\|_1 = \sum\limits_{i=1}^{n} |x_i|)$ the equilibration is unique but the convergence of the algorithm is slow (Sinkhorn (1964), (1967); Sinkhorn and Knopp).

7.3   Difficulties With Symmetric Matrices

If A is symmetric, then we allow $\beta^{-2} \mu_0$ instead of $\beta^{-1} \mu_0$ as the lower bound in the definitions of row (and column)-equilibrated. If A is symmetric, then A is row equilibrated iff A is column equilibrated iff A is equilibrated.

In order to equilibrate a symmetric matrix and still preserve the symmetry we seek a diagonal matrix D such that $D A D$ is equilibrated.

Let $A = \begin{bmatrix} 1 & 2 \\ 2 & 1/2 \end{bmatrix}$, $D_1 = \text{diag } \{1/2, 1\}$, $D_2 = \text{diag } \{1, 1/2\}$, and $D_3 = \text{diag } \{\sqrt{2}/4, \sqrt{2}\}$.

Then $D_1 A D_1 = \begin{bmatrix} 1/4 & 1 \\ 1 & 1/2 \end{bmatrix}$, $D_2 A D_2 = \begin{bmatrix} 1 & 1 \\ 1 & 1/8 \end{bmatrix}$, and

$D_3 A D_3 = \begin{bmatrix} 1/8 & 1 \\ 1 & 1 \end{bmatrix}$ are all equilibrated.

Criteria for choosing pivots are based on their size. It is an open problem whether all equilibrated forms of a symmetric matrix give satisfactory pivots as judged by the usual criteria.

There is no known algorithm for the symmetric case.

7.4 The Obvious Attempt

Let us consider $D A D$. Let $D = \text{diag } \{d_1,\ldots,d_n\}$ and A be an $n \times n$ symmetric matrix. Then $(DAD)_{ij} = d_i d_j A_{ij}$. Let us assume that no row of A is all zero.

The method which seems the more obvious is to equilibrated one row at a time: Let $D_1 = \text{diag}\{d_1,1,\ldots,1\}$.

Then $(D_1AD_1)_{ij} = \begin{cases} A_{ij} & \text{for } i,j > 1 \\ d_1 A_{ij} & \text{for } i = 1, j \neq 1 \\ & \text{or } j = 1, i \neq 1 \\ d_1^2 A_{11} & \text{for } i,j = 1 \end{cases}$

So choose $d_1^{-1} = \max \{\sqrt{|A_{11}|} , \max_{2 \leq j \leq n} |A_{1j}|\}$.

Now D_1AD_1 is symmetric and $\max_{1 \leq j \leq n} |(D_1AD_1)_{1j}| = 1$. However, when we try to equilibrate the second row we usually destroy the equilibration of the first row.

This method is related to the Sinkhorn algorithm which equilibrates in the one-norm, $\|x\|_1 = \sum_{i=1}^{n} |x_i|$, rather than the ∞-norm,

$\|x\|_\infty = \max_i |x_i|$, which we are using (Sinkhorn (1964), (1967); Sinkhorn and Knopp; Marcus and Newman).

7.5   Equilibration of Lower Triangular and Symmetric Matrices

The problem with the method in §7.4 is that we try to do too much at each step.

Let us consider $A = A^t = T + \Delta + T^t$, where Δ is a diagonal matrix and T is strictly lower triangular (i.e. $T_{ij} = 0$ if $j \geq i$).

Lemma 1: If $T + \Delta$ is row-equilibrated, then A is equilibrated.

Proof: Let us consider the ith row of A . Since $T + \Delta$ is row-equilibrated, $\max_{1 \leq j \leq i} |A_{ij}| = 1$. But $|A_{ji}| \leq 1$ for $j > i$, since $|T_{ij}| \leq 1$ and $A_{ji} = T_{ij}$ for $j > i$. Hence $\max_{1 \leq j \leq n} |A_{ij}| = 1$ for each i . $q.e.d.$

Now we shall show how to construct D such that $D(T + \Delta)D$ is row-equilibrated if $T + \Delta$ has no all zero row. In this case, by Lemma 1, $D A D$ is equilibrated.

In fact, it is eas   uilibrate the lower triangular part of a symmetric matrix and s.. ... preserve symmetry.

Lemma 2: Let B be an $n \times n$ lower triangular matrix (i.e. $B_{ij} = 0$ for $j > i$) with no all-zero row. For $1 \leq i \leq n$, let

$d_i^{-1} = \max \{ \sqrt{|B_{ii}|} , \max_{1 \leq j \leq i-1} |d_j B_{ij}| \}$. Then $D B D$ is row equilibrated, where $D = \text{diag} \{d_1, \ldots, d_n\}$.

__Proof__: By induction. Let $d_1 = 1/\sqrt{|B_{11}|}$, and $B_{11} \neq 0$ by hypothesis. So $|d_1^2 B_{11}| = 1$.

Assume that $d_1, \ldots, d_{i-1} > 0$ have been chosen so that $\max_{1 \leq k \leq j} |d_j d_k B_{jk}| = 1$ for $1 \leq j \leq i-1$.

Let $d_i^{-1} = \max \{ \sqrt{|B_{ii}|}, \max_{1 \leq j < i} |d_j B_{ij}| \}$. By hypothesis, $d_i^{-1} > 0$. Then $\max_j |d_i d_j B_{ij}| = 1$.

Hence the d_i exist for $1 \leq i \leq n$ by induction. Let $D = \text{diag} \{d_1, \ldots, d_n\}$. Then $D B D$ is row-equilibrated. $\quad q.e.d.$

7.6   The Algorithm for Null Rows in the Lower Triangle

But what do we do if $A_{11} = 0$ or if, for some i, $A_{ij} = 0$ for $1 \leq j \leq i$?

Let us form D as in §7.5 with the exception that we set $d_i = 1$ if $A_{ij} = 0$ for $1 \leq j \leq i$.

Then for all i, j: $|(D A D)_{ij}| \leq 1$.

$D A D$ fails to be equilibrated only if for some i:

(a) $A_{ij} = 0$ for $1 \leq j \leq i$, or

(b) $\max_{j > i} |d_j A_{ij}| < 1$.

If the ith row of A is not null, then the maximum in (b) is positive. For such i, define e_i by $e_i^{-1} = \max_{j > i} |d_j A_{ij}|$; for all other i, let $e_i = 1$. Let $E = \text{diag} \{e_1, \ldots, e_n\}$.

Theorem: Let A be an $n \times n$ symmetric matrix with no null row. Let D and E be constructed as above. Let $\Delta = D E$. Then Δ is a diagonal matrix, and $\Delta A \Delta$ is equilibrated.

Proof: If $e_i \neq 1$ then the maximal magnitude of row i in $D A D$ is raised to 1 by forming $E D A D E$, while in all other rows the ith element is increased in magnitude but not in excess of 1.

The theorem follows from §7.5. $q.e.d.$

7.7   Summary of Equilibration of Symmetric Matrices

If A is an $n \times n$ symmetric matrix with no null row, we can find a diagonal matrix D (in two sweeps, although only n steps) such that $D A D$ is equilibrated (in the ∞-norm, $\|x\|_\infty = \max_i |x_i|$).

We can express D by the following algorithm: For $i = 1(1)n$:

$$
\delta_i^{-1} =
\begin{cases}
\max \{ \sqrt{|A_{ii}|} , \max_{1 \leq j \leq i-1} |\delta_j A_{ij}| \} & \text{if } A_{ij} \neq 0 \text{ for} \\
& \text{some } j , 1 \leq j \leq i \\
1 & \text{if } A_{ij} = 0 \text{ for } 1 \leq j \leq i .
\end{cases}
$$

Then, for $i = 1(1)n$:

$$
d_i^{-1} =
\begin{cases}
\delta_i^{-1} & \text{if } A_{ij} \neq 0 \text{ for some } j , 1 \leq j \leq i \\
\max_{i+1 \leq j \leq n} |\delta_j A_{ji}| & \text{if } A_{ij} = 0 \text{ for } 1 \leq j \leq i
\end{cases}
$$

Let $D = \text{diag} \{d_1, \ldots, d_n\}$. Then $D A D$ is equilibrated.

The work required to equilibrate an $n \times n$ symmetric matrix with no all zero row is:

n	square roots
$n(n+1)$	multiplications .
$(n-1)^2$	additions

In practice the algorithm can be expressed in a very simple manner in Algol (see Appendix B) and can be performed in only n steps. (We would actually set $f_i = \delta_i^{-1} = 0$ if $A_{ij} = 0$ for $1 \leq j \leq i$ and then search for $\max\limits_{i+1 \leq j \leq n} |\delta_j A_{ji}|$ iff $f_i = 0$).

7.8 The Algorithm for Exponent Adjustment

In practice, we actually only require that $\beta^{-2} \leq \max\limits_{1 \leq j \leq n} |A_{ij}| \leq 1$

for every i instead of $\max\limits_{1 \leq j \leq n} |A_{ij}| = 1$ for every i so that we need only adjust the exponents of the elements of A .

Then our algorithm takes the following form:

Let $A_{ij} = \gamma_{ij} \beta^{\alpha_{ij}}$ for $j \neq i$, where $\beta^{-1} \leq |\gamma_{ij}| \leq 1$, unless $A_{ij} = 0$ when we take $\gamma_{ij} = 0 = \alpha_{ij}$.

Let $A_{ii} = \gamma_{ii} \beta^{\alpha_{ii}}$, where $\beta^{-2} \leq |\gamma_{ii}| \leq 1$, unless $A_{ii} = 0$ when we take $\gamma_{ii} = 0 = \alpha_{ii}$.

For $i = 1(1)n$: set $\delta_i = \begin{cases} 0 & \text{if } A_{ij} = 0 \text{ for } 1 \leq j \leq i \\ \max\{\tfrac{1}{2}\alpha_{ii}, \max\limits_{1 \leq j < i} (\alpha_{ij} - \delta_j)\} & \text{otherwise} \end{cases}$

For $i = 1(1)n$: set $d_i = \begin{cases} \delta_i & \text{if } \delta_i \neq 0 \\ \max\limits_{i+1 \leq j \leq n} (\alpha_{ji} - \delta_j) & \text{if } \delta_i = 0 \end{cases}$

This will give $\beta^{-2} \le \max_{1 \le j \le n} |\beta^{d_i + d_j} A_{ij}| \le 1$ for every i.

(This equilibration can be performed very rapidly in machine language.)

Chapter 8 : <u>Unequilibrated Diagonal Pivoting</u>

8.1 <u>Maximal Off-diagonal Element</u>

In order to obtain a lower bound of $\mu_0^2 - \mu_1^2$ for ν_p in Theorem 1 in §6.3, we needed the fact that, due to equilibration, there existed an element of maximal absolute value in the first column. However, if $\mu_1 < \mu_0$, then there exist integers i,j with $i > j$, such that $|A_{ij}| = \mu_0$. We need only bring the element A_{ij} up to the $(2,1)$ position and then we will have a 2×2 pivot with a maximal off-diagonal element. We shall call this variation unequilibrated diagonal pivoting.

Let $\mu_0 = \max_{i,j} |A_{ij}| = |A_{rs}|$, where r,s are the least such integers. Let $\mu_1 = \max_i |A_{ii}|$.

Let $\nu_b = |A_{rr} A_{ss} - A_{rs}^2|$.

This strategy involves:

(1) finding μ_1 and the least integer k with $|A_{kk}| = \mu_1$;

(2) finding μ_0 and the least integers r,s with $|A_{rs}| = \mu_0$;

(3) choosing a 1×1 or 2×2 pivot according to S_α (§5.7);

(4) for a 1×1, interchanging rows and columns 1 with k so that $|P| = \mu_1$ (§5.3);

(5) for a 2×2, interchanging rows and columns 1 with r and 2 with s so that $|\det P| = \nu_b$ and $|A_{21}| = \mu_0$ (§5.4).

This procedure is repeated for each reduced matrix.

Note that calculating ν_b requires only 2 multiplications instead of $n(n-1)$ for ν_c and $2(n-1)$ for ν_p.

Clearly from the definitions of ν_b, ν_p, and ν_c and from Lemma 3 of §5.5, we have:

Lemma 1: $\nu_b \leq \nu_p \leq \nu_c \leq \mu_0^2 + \mu_1^2$.

8.2 Bounding ν_b

Let us now bound ν_b from below. From §8.1, we may assume $\nu_b = |A_{11} A_{22} - A_{21}^2|$ and $|A_{21}| = \mu_0$.

Theorem 1: If $|A_{21}| = \mu_0$, then $\mu_0^2 - \mu_1^2 \leq \nu_b \leq \mu_0^2 + \mu_1^2$.

Proof: The upper bound follows from Lemma 1.

Since $|A_{21}| = \mu_0$, $\nu_b = |A_{11} A_{22} - A_{21}^2| = \mu_0^2 - A_{11} A_{22} \geq \mu_0^2 - \mu_1^2$. $q.e.d.$

Here, as in §6.2 and §6.5, symmetry was used to get the lower bound on ν_b. By symmetry, if $\mu_0 = |A_{21}|$ then $|A_{12}| = \mu_0$.

If A were not symmetric, then $\mu_0 = |A_{21}|$ does not imply that $|A_{12}| = \mu_0$ (in fact we could have $A_{12} = 0$). Thus no such non-negative lower bound on the determinant of 2×2 submatrices can exist for non-symmetric matrices.

Thus Theorem 1 implies that Lemma 4 in §5.5 holds for ν_b if $\mu_1 < \mu_0$. According to S_α (§5.7), we would choose a 1×1 pivot for the reduced matrix $A^{(k)}$ of order k iff $\mu_1^{(k)} \geq \alpha_0 \mu_0^{(k)}$, where $\alpha_0 = (1 + \sqrt{17})/8$.

In Chapters 10-12, we shall see that stability of this strategy (for unequilibrated matrices) follows from the above.

8.3   Comments on this Strategy

From §8.1, we see that we need do no searching over the 2×2 principal minors, but we merely choose that principal minor with maximal off-diagonal element. In Chapter 9 we shall see that this searching for the $\mu_0^{(k)}$ requires between $\sim \frac{1}{12} n^3$ and $\sim \frac{1}{6} n^3$ additions and no multiplications.

We used the terms "complete" and "partial" strategies in Chapter 4 to distinguish between searching over all the 2×2 principal minors and over the 2×2 principal minors with off-diagonal element in the first column.

In analogy with Gaussian elimination with complete pivoting, we would like to call our strategy in §8.1 a complete pivoting strategy since we search all of $A^{(k)}$ for its maximal element, but we do not wish to cause confusion with the use of the word "complete" in Chapter 4 where it meant searching all the 2×2 principal minors.

Now, in analogy with Gaussian elimination with partial pivoting, we ask if there could be a partial strategy where we search only the first column of $A^{(k)}$ for its maximal element.

Such a partial strategy requires at most only $\frac{1}{2} n (n-1)$ additions to calculate the maximal element in the first column of all the reduced matrices. If such a partial strategy were stable, then this strategy would require only $\sim \frac{1}{6} n^3$ multiplications and $\sim \frac{1}{6} n^3$ additions to solve $A x = b$, $A = A^t$, $\det A \neq 0$.

However, any such partial strategy is unstable for unequilibrated matrices, as the following example shows:

$$
A = \begin{bmatrix} \frac{1}{2}\alpha\,\epsilon & \epsilon & \epsilon \\ \epsilon & \frac{1}{2}\alpha\,\epsilon & 1 \\ \epsilon & 1 & \frac{1}{2}\alpha\,\epsilon \end{bmatrix}
$$

where $0 < \alpha < 1$ and $0 < \epsilon \ll 1$.

8.4 A Partial Strategy for Equilibrated Reduced Matrices

Clearly, if $A = A^{(n)}$ and each of its reduced matrices $A^{(k)}$ is equilibrated (i.e. $\max_{i,j} |A_{ij}^{(k)}| = \mu_0^{(k)}$ for each i), then the partial strategy, whereby we choose the 2×2 principal submatrix whose off-diagonal element is the maximal element in the first column of $A^{(k)}$, is stable since, by the equilibration, such a maximal element in the first column of $A^{(k)}$ is also a maximal element of $A^{(k)}$.

But we would have to equilibrate A at the start, and then equilibrate each reduced matrix $A^{(k)}$.

Let us now consider such a partial strategy when A is equilibrated, but we do not equilibrate the reduced matrices.

8.5 A Partial Strategy for Unequilibrated Reduced Matrices

Let A be an n×n symmetric equilibrated non-singular matrix with $\max_j |A_{ij}| = \mu_0$ for each i . We shall now consider the partial strategy defined in §§8.3-8.4, but we shall not equilibrate the reduced matrices $A^{(k)}$ for $k < n$.

Let $A^{(k)}$ be the reduced matrix of order k ; let $A^{(n)} = A$.

Let $\mu_0^{(k)} = \max_{i,j} |A_{ij}^{(k)}|$. (We shall not actually calculate $\mu_0^{(k)}$).

Let $\mu_1^{(k)} = \max_i |A_{ii}^{(k)}| = |A_{jj}^{(k)}|$ say. We shall assume we have

interchanged rows and columns so that $|A_{11}^{(k)}| = \mu_1^{(k)}$. Then let

$\lambda^{(k)} = \max_i |A_{11}^{(k)}|$. So $\lambda^{(k)} \leq \mu_0^{(k)}$, while $\lambda^{(n)} = \mu_0^{(n)} = \mu_0$.

We shall use a 1×1 pivot iff $\mu_1^{(k)} \geq \alpha \lambda^{(k)}$. Let m be any
multiplier (§5.1).

If $\mu_1^{(k)} \geq \alpha \lambda^{(k)}$ then we use $A_{11}^{(k)}$ as a 1×1 pivot, and

$|m| \leq \lambda^{(k)}/\mu_1^{(k)} \leq 1/\alpha$, while $\max_{i,j} |A_{ij}^{(k-1)}| \leq \mu_0^{(k)} + \lambda^{(k)}/\alpha \leq$

$\mu_0^{(k)} (1 + 1/\alpha)$.

If $\mu_1^{(k)} < \alpha \lambda^{(k)}$ then we interchange so that $|A_{21}^{(k)}| = \lambda^{(k)}$.

Let $\nu^{(k)} = |A_{11}^{(k)} A_{22}^{(k)} - A_{21}^{(k)2}|$. So $\nu^{(k)} \geq \lambda^{(k)2} - \mu_1^{(k)2} \geq (1 - \alpha^2) \lambda^{(k)2}$.

Then $|m| \leq \lambda^{(k)} (\mu_0^{(k)} + \mu_1^{(k)})/\nu^{(k)}$ and

$\max_{i,j} |A_{ij}^{(k-2)}| \leq \mu_0^{(k)} + 2\lambda^{(k)2} (\mu_0^{(k)} + \mu_1^{(k)})/\nu^{(k)} < [1 + 2/(1 - \alpha)] \mu_0^{(k)}$.

As in §5.7 we would choose $\alpha = \alpha_0 = (1 + \sqrt{17})/8$. Then

$\max_{i,j} |A_{ij}^{(k)}| < (2.57)^{n-k} \mu_0$ for each $A^{(k)}$.

But we cannot obtain a bound on $A^{(k)}$ as in Chapters 11-12 since
we cannot express bounds on $\nu^{(k)}$ in terms of $\mu_0^{(k)}$.

We leave the significance of this method (§8.5) in relation to the method in §§8.1-8.4 as an open question.

Chapter 9 : Operation Count

9.1 Solution by Diagonal Pivoting

We now consider the amount of work required to solve $AX = B$ by the (unequilibrated) diagonal pivoting method (Chapter 8), where A is an $n \times n$ non-singular symmetric matrix and B is an $n \times k$ matrix, i.e. there are k right hand sides.

In matrix notation, we perform the following steps:

(1) $A = M D M^t$.

(2) $C = M^{-1} B$.

(3) $Y = D^{-1} C$.

(4) $X = M^{-t} Y$.

Here M^{-t} means $(M^{-1})^t$.

Let p be the number of 1×1 pivots used. So $q = \frac{1}{2}(n - p)$ pivots of order 2 are used. Let $A^{(i)}$ be the reduced matrix of order i .

Definition:

$$\text{pivot } [i] = \begin{cases} 1 & \text{if } A^{(i)} \text{ uses a } 1 \times 1 \text{ pivot} \\ 2 & \text{if } A^{(i)} \text{ uses a } 2 \times 2 \text{ pivot} \\ 0 & \text{if } A^{(i+1)} \text{ uses a } 2 \times 2 \text{ pivot} \end{cases}$$

We shall use the term Mults (Adds) to mean the number of multiplications (additions). We shall count a comparison as an addition.

We shall use $\Sigma^{(j)}$ to denote summation over those indices i , $1 \le i \le n$, such that pivot$[i] = j$.

9.2 Summary of the Work Required

Steps (1), (2), (3), (4) in §9.1 require:

$$\text{Mults} = \frac{1}{6}n^3 + (k + \frac{1}{2})n^2 + (\frac{11}{4} - 2k)n + (k + 3/2)p + \sum^{(2)} 3i$$

$$\leq \frac{1}{6}n^3 + (k + \frac{5}{4})n^2 + (\frac{4}{3} - 2k)n + \frac{1}{12}(2k - 7)^2 \sim \frac{1}{6}n^3$$

$$\text{Adds} = \frac{1}{4}n^3 + (k + 5/8)n^2 + (\frac{3}{4} - 2k)n + (k - 3/8)p + \sum^{(1)} \frac{1}{4}i\ (i - 1)$$

$$\sim \frac{1}{4}n^3 + \frac{1}{12}p^3$$

$$\leq \frac{1}{3}n^3 + (k + \frac{1}{2})n^2 - (k - \frac{1}{6})n \sim \frac{1}{3}n^3\ .$$

For $k = 1$ (i.e. one right hand side), let us compare the work required for the diagonal pivoting method (applicable to all non-singular symmetric matrices) with that required for Cholesky's method (applicable only to positive-definite matrices).

| | Cholesky | Diagonal Pivoting | |
	Exact	Lower Bound	Upper Bound
Mults	$\frac{1}{6}n^3 + \frac{3}{2}n^2 + \frac{1}{3}n$	$\frac{1}{6}n^3 + \frac{3}{2}n^2 + \frac{4}{3}n$	$\frac{1}{6}n^3 + \frac{9}{4}n^2 - \frac{2}{3}n + \frac{25}{12}$
Adds	$\frac{1}{6}n^3 + n^2 - \frac{7}{6}n$	$\frac{1}{4}n^3 + \frac{13}{8}n^2 - \frac{5}{4}n$	$\frac{1}{3}n^3 + \frac{3}{2}n^2 - \frac{5}{6}n$
Root Recipro-cals	n	0	0

Since the time required for a computer to perform a multiplication is much longer than for an addition, the requirement of $\sim \frac{1}{6}n^3$ multiplications and between $\sim \frac{1}{4}n^3$ and $\sim \frac{1}{3}n^3$ additions is very satisfactory for symmetric indefinite matrices.

If A is positive definite, then Cholesky's method is preferable to the diagonal pivoting method. See also Appendix A (§A.1).

9.3 Forming (1) A = M D M^t

First we decompose A into A = M D M^t .

Let us consider the reduced matrix $A^{(i)}$ of order i .

The following chart shows our course of action.

find $\mu_1^{(i)}$: Adds = i - 1

find $\mu_0^{(i)}$: Adds = $\frac{1}{2}$i (i-1)

$\mu_1^{(i)} \geq \alpha \, \mu_0^{(i)}$? : Mults, Adds = 1

yes   no

use 1×1   use 2×2

interchange   interchange

Mults = i-1: calculate   calculate det:{ Mults = 2,
              multipliers   Adds = 1

Mults, Adds }   calculate   :{ Mults = 6(i-2),
= $\frac{1}{2}$ i (i-1)   :form $A^{(i-1)}$ multipliers   Adds = 2(i-2)

                                           form $A^{(i-2)}$:{ Mults, Adds
                                                                 =(i-1)(i-2)

i := i-1   i := i-2
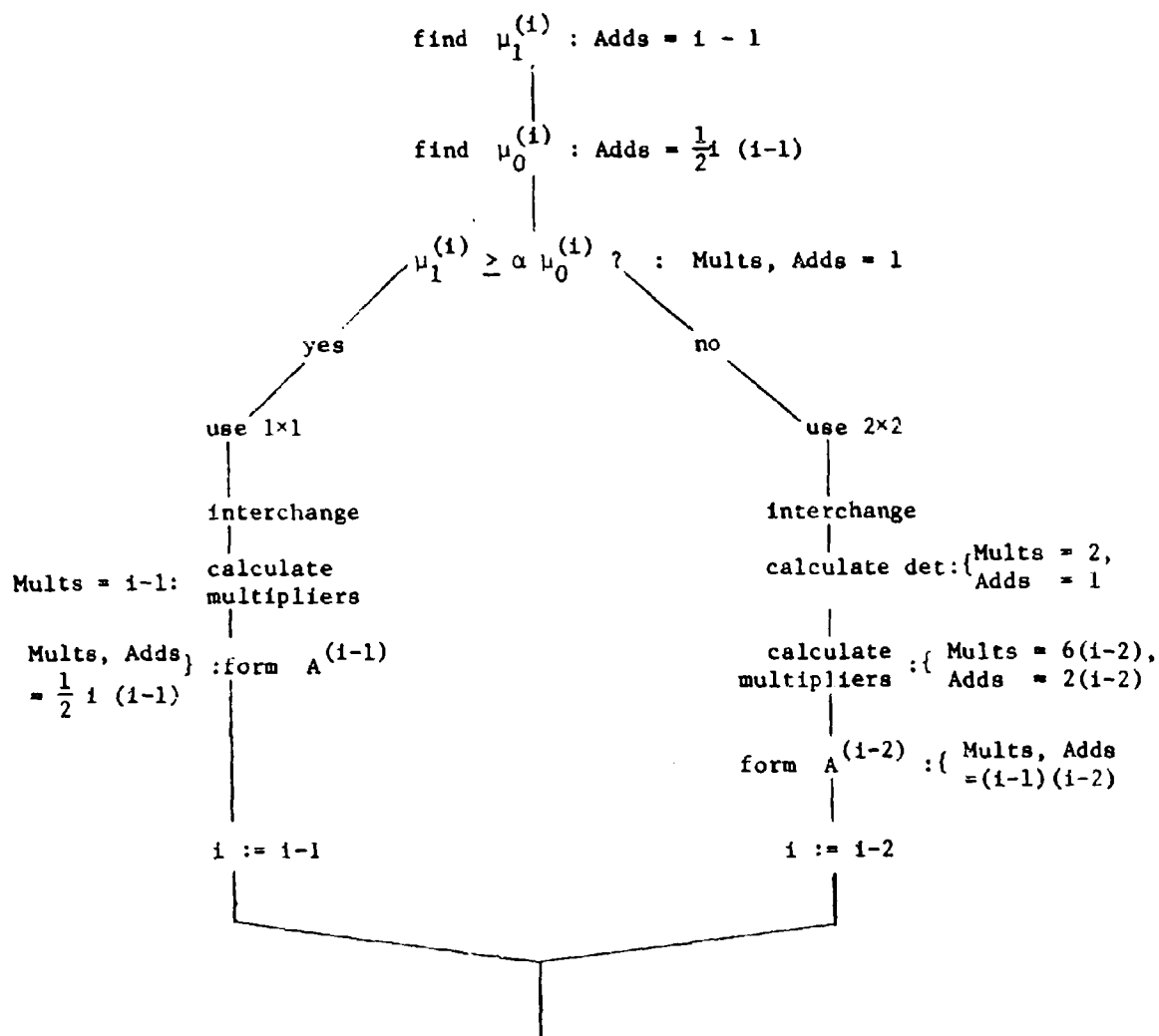
Thus for $A^{(i)}$:

Mults $= \frac{1}{2} i(i+1)$ and Adds $= i^2$ if pivot[i] $= 1$,

Mults $= i^2 + 3i - 7$ and Adds $= (i-1)(\frac{3}{2} i+1)$ if pivot[i] $= 2$.

Recall that there are p 1×1 pivots and $q = \frac{1}{2}(n-p)$ 2×2 pivots.

For step (2):

Mults $= \sum^{(1)} \frac{1}{2} i (i+1) + \sum^{(2)} (i^2 + 3i - 7)$

$= \sum^{(1)} \frac{1}{2} i (i+1) + \sum^{(2)} \{\frac{1}{2} i (i+1) + \frac{1}{2} (i-1)i + 3i - 7\}$

$= \sum \frac{1}{2} i (i+1) + \sum^{(2)} (3i - 7)$,

$(9.3.1) \quad = \frac{1}{6} n^3 + \frac{1}{2} n^2 - \frac{19}{6} n + \frac{7}{2} p + 3 \sum^{(2)} i$

$\geq \frac{1}{6} n^3 + \frac{1}{2} n^2 + \frac{1}{3} n$

Adds $= \sum^{(1)} i^2 + \sum^{(2)} (i-1)(\frac{3}{2} i+1)$

$= \sum^{(1)} \{(\frac{3}{4} i^2 + \frac{1}{2} i) + (\frac{1}{4} i^2 - \frac{1}{2} i)\} +$

$\sum^{(2)} \{\frac{3}{4} i^2 + \frac{1}{2} i + \frac{3}{4} (i-1)^2 + \frac{1}{2} (i-1) - \frac{5}{4}\}$

$= \sum_{i=1}^{n} (\frac{3}{4} i^2 + \frac{1}{2} i) - \frac{5}{8} (n-p) + \sum^{(1)} \frac{1}{4} i (i-2)$

$(9.3.2) \quad = \frac{1}{4} n^3 + \frac{5}{8} n^2 - \frac{1}{4} n + \frac{5}{8} p + \frac{1}{4} \sum^{(1)} i (i-2)$

$\geq \frac{1}{4} n^3 + \frac{5}{8} n^2 - \frac{1}{4} n$.

Now let us bound $3 \sum^{(2)} i$ and $\frac{1}{4} \sum^{(1)} i (i-2)$ from above.

$\sum^{(2)} i \leq \sum_{j=1}^{q} (p+2j) = \frac{1}{4} (n^2-p^2) + \frac{1}{2} (n-p)$ and

$$\sum^{(1)} i \ (i-2) \leq \sum_{i=n-p+1}^{n} i \ (i-2) = p\{n^2-(p+1)n + \tfrac{1}{3} p^2 + \tfrac{1}{2} p - \tfrac{5}{6} \}$$

So we have the following upper bounds for (1):

(9.3.3)   Mults $\leq \tfrac{i}{6} n^3 + \tfrac{5}{4} n^2 - \tfrac{5}{3} n - \tfrac{3}{4} p^2 - \tfrac{3}{2} p$.

(9.3.4)   Adds $\leq \tfrac{1}{4} n^3 + \tfrac{1}{12} p^3 + (\tfrac{1}{4} p + \tfrac{5}{8}) n^2 - \tfrac{1}{4} (p^2+p+1)n + \tfrac{1}{8} p^2 + \tfrac{5}{12} p$.

9.4   Solving (2), (3), (4)

For (2) $C = M^{-1} B$:

Mults, Adds $= (\tfrac{1}{2} n^2 - n + \tfrac{1}{2} p) k$.

(Note that $M_{i+1,i} = 0$ if pivot [i] = 2.)

For (3) $Y = D^{-1} C$:

Mults $= 3n - 2p$, Adds $= n - p$.

For (4) $X = M^{-t} Y$:

Mults, Adds $= (\tfrac{1}{2} n^2 - n + \tfrac{1}{2} p) k$.

9.5   Total Work Required

Thus steps (1), (2), (3), (4) together require:

Mults $= \tfrac{1}{6} n^3 + (k + \tfrac{1}{2}) n^2 + (\tfrac{11}{4} - 2k) n + (k + \tfrac{3}{2}) p + 3 \sum^{(2)} i$,

Mults $\leq \tfrac{1}{6} n^3 + (k + \tfrac{5}{4}) n^2 + (\tfrac{4}{3} - 2k) n + (k - \tfrac{7}{2}) p - \tfrac{3}{4} p^2$,

Mults $\geq \tfrac{1}{6} n^3 + (k + \tfrac{1}{2}) n^2 + (\tfrac{10}{3} - 2k) n$.

Adds $= \tfrac{1}{4} n^3 + (k + \tfrac{5}{8}) n^2 + (\tfrac{3}{4} - 2k) n + (k - \tfrac{3}{8}) p + \tfrac{1}{4} \sum^{(1)} i \ (i-2)$,

Adds $\leq \tfrac{1}{4} n^3 + (k + \tfrac{5}{8}) n^2 + (\tfrac{3}{4} - 2k) n + \tfrac{1}{12} p^3 + (\tfrac{1}{8} - \tfrac{n}{4}) p^2 +$

$(\tfrac{1}{4} n - \tfrac{1}{4} n + k - \tfrac{7}{12}) p$,

Adds $\geq \tfrac{1}{4} n^3 + (k + \tfrac{5}{8}) n^2 + (\tfrac{3}{4} - 2k) n$.

9.6 Upper Bound on Mults

We would like a minimal upper bound on Mults that is independent of p, where $0 \le p \le n$.

Let $f(x) = (k - \frac{7}{2})x - \frac{3}{4}x^2$. By elementary calculus,

$f(p) \le f(\frac{2k - 7}{3}) = \frac{1}{12}(2k - 7)^2$. Thus Mults $\le \frac{1}{6}n^3 + (k + \frac{5}{4})n^2 +$

$(\frac{4}{3} - 2k)n + \frac{1}{12}(2k-7)^2 = \frac{1}{6}n^3 + \frac{9}{4}n^2 - \frac{2}{3}n + \frac{25}{12}$ (when $k = 1$).

9.7 Upper Bound on Adds

Now we would also like a minimal upper bound on Adds that is independent of p.

Let $g(x) = \frac{1}{12}x^3 + (\frac{1}{8} - \frac{n}{4})x^2 + (\frac{1}{4}n - \frac{1}{4}n + k - \frac{7}{12})x$.

Since $g'(x) > 0$ on $[0,n]$ we have

$g(p) \le g(n) = \frac{1}{12}n^3 - \frac{1}{8}n^2 + (k - \frac{7}{12})n$. Thus

Adds $\le \frac{1}{3}n^3 + (k + \frac{1}{2})n^2 - (k - \frac{1}{6})n$

$= \frac{1}{3}n^3 + \frac{3}{2}n^2 - \frac{5}{6}n$ (where $k = 1$).

Chapter 10 : <u>Error Analysis for Diagonal Pivoting</u>

10.1   <u>Introduction</u>

Let us attempt to solve $A x = b$ by the diagonal pivoting
method. If x_c is the solution we obtain from the computer, we may
consider x_c as the exact solution of the system $(A + E) z = b$.
As in §2.5, we are interested in showing that the elements of E are
small in comparison to the corresponding elements of A (Wilkinson,
1960, 1961, 1963, 1965). Such an analysis of E is called a back-
ward error analysis.

10.2   <u>The Occurrences of Error</u>

Suppose that we could perform the diagonal pivoting method in
exact arithmetic. In order to solve $A x = b$, we perform the fol-
lowing steps (see §9.1):

(1) $A = M D M^t$ (the decomposition)

(2) $c = M^{-1} b$ (the new right hand side)

(3) $y = D^{-1} c$ (solve the 1×1 and 2×2 systems)

(4) $x = M^{-t} y$ (recover the solution)

However, in finite precision arithmetic, we have error at each
step. Instead of decomposing A into $M D M^t$, we obtain M and D
such that $M D M^t = A + F$. Instead of calculating $M^{-1} b$, $D^{-1} c$,
$M^{-t} y$, we actually obtain $c = (M + M_1)^{-1} b$, $y = (D + \delta D)^{-1} c$,
$x = (M + M_2)^{-t} y$ for some perturbations M_1 , δD , M_2 respectively.
Thus we actually perform:

$$(1) \quad A + F = M \, D \, M^t$$

$$(2) \quad c = (M + M_1)^{-1} \, b$$

$$(3) \quad y = (D + \delta D)^{-1} \, c$$

$$(4) \quad x = (M + M_2)^{-t} \, y$$

10.3   The Error Matrix E

Thus we have $\quad b = (M + M_1) \, c = (M + M_1)(D + \delta D) \, y$

$= (M + M_1)(D + \delta D)(M + M_2)^t \, x = \{M \, D \, M^t + M_1 \, (D + \delta D)(M + M_2)^t$

$+ \, M \, [\delta D \, (M + M_2)^t + D \, M_2^t \,]\} \, x$. But $M \, D \, M^t = A + F$, while

$(A + E) \, x = b$. Hence

$$E = F + M_1 \, (D + \delta D)(M + M_2)^t + M \, [\delta D (M + M_2)^t + D \, M_2^t \,] \, .$$

If we can bound the elements of F , M , D , M_1 , M_2 , and δD , then we can bound the elements of E . We shall see that most of the error lies in F , in other words, most of the error occurs when obtaining the decomposition $M \, D \, M^t$ of A .

10.4   Notation

In the following sections, the symbol ϵ, η will stand for any numbers with $|\epsilon| \leq 2^{-t}$, $|\eta| \leq (1.06) \, 2^{-2t}$, where t is the number of binary digits in the computer. Each occurrence of ϵ or η in an equation should be indexed, but we shall suppress these indices for the sake of clarity.

$\sum_i'$ and $\sum_i''$ shall denote the summation over those indices k , $i \leq k \leq n$, with pivot[k] = 1 and 2 respectively.

We shall write $g(t) = 0 (2^{-t})$ if $\lim_{t \to \infty} 2^t g(t)$ is finite.

10.5   Summary of the Error Analysis

In the following sections we shall show:

(10.5.1) $|F_{ij}| \, 2^t < [1 + 3.01/\alpha] \sum_1' \mu_0^{(k)} + [1 + 11.02/(1-\alpha)] \sum_1'' \mu_0^{(k)}$

for $i \geq j$, and $F = F^t$.

(10.5.2) $|E_{ij}| \leq |F_{ij}| + \epsilon$, where $\epsilon < 2^{-t} (1+\alpha) \max_k \mu_0^{(k)} n^2 \times$

$\max \{1/\alpha^2, 1/(1-\alpha^2)\}$.

Also, for $\alpha = \alpha_0$, we shall show:

(10.5.3) $|F_{ij}| \, 2^t < 5.71 \sum_1' \mu_0^{(k)} + 31.6 \sum_1'' \mu_0^{(k)}$ for $i \geq j$.

(10.5.4) $\|E\| < 2^{-t} \max_k \mu_0^{(k)} (23.54) n^2$

where $\| \ \|$ is the one-norm: $\|E\| = \max_j \sum_{i=1}^{n} |E_{ij}|$.

In Chapters 11 and 12, we shall show for $\alpha = \alpha_0$:

(10.5.5) $\max_{i,j} |F_{ij}| < (15.8) n \, 2^{-t} \sqrt{n} \, f(n) \, \mu_0 \, (3.07) \, n^{0.446}$,

(10.5.6) $\|F\| < (15.8) n^2 \, 2^{-t} \sqrt{n} \, f(n) \, \mu_0 \, (3.07) \, n^{0.446}$,

(10.5.7) $\|E\| < (23.54) n^2 \, 2^{-t} \sqrt{n} \, f(n) \, \mu_0 \, (3.07) n^{0.446}$.

For Gaussian elimination with complete pivoting (see §2.5):

(10.5.8) $\|F\| \leq (2.01) n^2 \, 2^{-t} \sqrt{n} \, f(n) \, \mu_0$, where $L U = A + F$.

For further remarks, see §10.12 and §10.16.

10.6 The Decomposition for the Reduced Matrix $A^{(r)}$

Let $A^{(r)}$ be the reduced matrix of order r, $1 \leq r \leq n$. Let $s = \text{pivot}[r]$ (see §9.1). Thus $s = 1$ or 2.

Let $A_s^{(r)}$ be the matrix resulting from deleting the first s rows and columns from $A^{(r)}$.

Let $A^{(r)} = \begin{bmatrix} P_s^{(r)} & C_s^{(r)\,t} \\ \hline C_s^{(r)} & A_s^{(r)} \end{bmatrix}$, where $P_s^{(r)}$ is $s \times s$.

$C_s^{(r)}$ is $(r-s) \times s$. We shall use $P_s^{(r)}$ as the $s \times s$ pivot.

Let $M_s^{(r)}$ be the $(r-s) \times s$ matrix resulting from calculating $C_s^{(r)} (P_s^{(r)})^{-1}$ in finite precision. Let $\Theta_s^{(r)}$ be the $(r-s) \times s$ error matrix : $\Theta_s^{(r)} = M_s^{(r)} - C_s^{(r)} P_s^{(r)-1}$. Let $A^{(r-s)}$ be the reduced matrix of order $r-s$ resulting from calculating $A_s^{(r)} - M_s^{(r)} C_s^{(r)t}$ in finite precision. Let $G^{(r-s)}$ be the $(r-s) \times (r-s)$ error matrix :

$$G^{(r-s)} = A^{(r-s)} - A_s^{(r)} + M_s^{(r)} C_s^{(r)t} .$$

But $M_s^{(r)} C_s^{(r)t} = M_s^{(r)} [(M_s^{(r)} - \Theta_s^{(r)}) P_s^{(r)}]^t =$

$$M_s^{(r)} P_s^{(r)} M_s^{(r)t} - M_s^{(r)} P_s^{(r)} \Theta_s^{(r)t} .$$

Hence $A_s^{(r)} + F^{(r-s)} = M_s^{(r)} P_s^{(r)} M_s^{(r)t} + A^{(r-s)}$, where

$$F^{(r-s)} = G^{(r-s)} + M_s^{(r)} P_s^{(r)} \Theta_s^{(r)t} .$$

10.7 The Error Matrix F for the Decomposition

From §10.6, we have $A + F = M D M^t$, where $F = F^t$ and for $i \geq j$: $F_{ij} = G_{ij} + (M D \Theta^t)_{ij}$, where for $i \geq j$:

$$G_{ij} = \sum_{k=2}^{j} \varepsilon_{ij}^{(k)} \quad \text{with}$$

$$\varepsilon_{ij}^{(k)} = \begin{cases} G_{i-k,j-k}^{(n-k)} & \text{if pivot}[n-k+1] = 1 \\[2ex] G_{i-k-1,j-k-1}^{(n-k-1)} & \text{if pivot}[n-k+1] = 2 \\[2ex] 0 & \text{if pivot}[n-k+1] = 0 \end{cases} ,$$

$$\text{and for } j \geq k : \quad O_{jk} = \begin{cases} (\Theta_1^{(n-k)})_{j-k,1} & \text{if pivot}[n-k+1] = 1 \\[2ex] (\Theta_2^{(n-k-1)})_{j-k-1,1} & \text{if pivot}[n-k+1] = 2 \\[2ex] (\Theta_2^{(n-k)})_{j-k-1,2} & \text{if pivot}[n-k+1] = 0 \end{cases} .$$

10.8   The Error Matrices M_1 , δD, M_2

$(M + M_1) c = b$. From §10.6,

$$M_{ij} = \begin{cases} (M_1^{(n-j)})_{i-j,1} & \text{if pivot}[n-j+1] = 1 \\[2ex] (M_2^{(n-j-1)})_{i-j-1,1} & \text{if pivot}[n-j+1] = 2 \\[2ex] (M_2^{(n-j)})_{i-j-1,2} & \text{if pivot}[n-j+1] = 0 \end{cases}$$

for $i > j$, with $M_{j+1,j} = 0$ if pivot$[n-j+1] = 2$.

while $M_{jj} = 1$ for every j , and $M_{ij} = 0$ for $j > i$.

Thus $(M_1)_{ij} = 0$ for $j \geq i$ and $(M_1)_{j+1,j} = 0$ if

pivot$[n-j+1] = 2$.

$(D + \delta D)y = c$. From §10.6, D is a block diagonal matrix with

blocks of order 1 and 2. The blocks are the pivots $P_s^{(r)}$ in §10.6,

s = pivot[r] . Also, δD has the same block structure as D .

$(M + M_2)^t x = y$. M_2^t has the same structure as M_1 .

10.9   Floating Point Error Analysis

Since most computers now perform calculations in floating point arithmetic rather than in fixed point (see Wilkinson, 1965, pp. 110–188), we shall give an error analysis only in floating point.

Then, from Wilkinson (1965), pp. 114–117, we have $z = f\ell(x * y) \equiv (x * y)(1 + \varepsilon)$, where $|\varepsilon| \le 2^{-t}$, t is the number of binary digits used by the computer, and $*$ is any arithmetic operation $(+, -, \times, \div)$.

Further, we shall assume that the computer can accumulate inner-products in floating point arithmetic. Then $\left| f\ell(x_1 y_1 + \ldots + x_n y_n) - (x_1 y_1 + \ldots + x_n y_n) \right| \le \sum_{i=1}^{n} i |\gamma_i x_i y_i|$, where $|\gamma_i| < \frac{3}{2}(1.06) 2^{-2t}$. As in §10.4, we shall assume η is any number such that $|\eta| \le (1.06) 2^{-2t}$ and we shall suppress indices. From §10.4, we may write $|\eta| \le 2^{-t} 0(2^{-t})$.

10.10   Floating Point Analysis for F

Consider $A^{(r)}$. $|A_{ij}^{(r)}| \le \mu_0^{(r)}$. Let $i \ge j$.

Case 1: s = pivot[r] = 1 .

$M_1^{(r)}$, $O_1^{(r)}$, and $C_1^{(r)}$ are $(r-1) \times 1$. $P_s^{(r)} = A_{11}^{(r)} = \mu_1^{(r)}$.

So $A_{ij}^{(r-1)} = [(A_1^{(r)})_{ij} - (M_1^{(r)})_i (C_1^{(r)})_j (1 + \varepsilon)] (1 + \varepsilon)$.

So $E_{ij}^{(r-1)} = (A_1^{(r)})_{ij}\epsilon - (M_1^{(r)})_1 (C_1^{(r)})_j \epsilon (2 + \epsilon)$.

$|(A_1^{(r)})_{ij}| = |A_{i+1,j+1}^{(r)}| \leq \mu_0^{(r)}$. $|(C_1^{(r)})_j| = |A_{j+1,1}^{(r)}| \leq \mu_0^{(r)}$.

Thus $|E_{ij}^{(r-1)}| \leq 2^{-t} \mu_0^{(r)} \{1 + (2+2^{-t}) |M_1^{(r)}|_1 \}$.

$(M_1^{(r)})_j = A_{j+1,1}^{(r)} A_{11}^{(r)-1} (1 + \epsilon)$

$\qquad = (C_1^{(r)} P_1^{(r)-1})_j (1 + \epsilon)$. So $(\Theta_1^{(r)})_j = \epsilon(C_1^{(r)} P_1^{(r)-1})_j$.

$|(M_1^{(r)})_j P_1^{(r)}| \leq \mu_0^{(r)} (1 + 2^{-t})$. $|(C_1^{(r)} P_1^{(r)-1})_j| = |A_{j+1,1}^{(r)} A_{11}^{(r)-1}| \leq$

$\qquad \mu_0^{(r)}/\mu_1^{(r)} \leq 1/\alpha$.

So $|(\Theta_1^{(r)})_j| \leq 2^{-t}/\alpha$ and $|(M_1^{(r)})_j| \leq (1 + 2^{-t})/\alpha$. Thus

(10.10.1) $|E_{ij}^{(r-1)}| \leq 2^{-t} \mu_0^{(r)} [1 + 2/\alpha + O(2^{-t})]$, and

(10.10.2) $|(M_1^{(r)} P_1^{(r)} \Theta_1^{(r)t})_{ij}| \leq \mu_0^{(r)} (1 + 2^{-t}) 2^{-t}/\alpha =$

$\qquad \mu_0^{(r)} 2^{-t} [1/\alpha + O(2^{-t})]$.

Case 2: $s = \text{pivot}[r] = 2$.

$M_2^{(r)}, \Theta_2^{(r)}$, and $C_2^{(r)}$ are $(r-2)\times 2$. $P_2^{(r)} = \begin{bmatrix} A_{11}^{(r)} & A_{21}^{(r)} \\ A_{21}^{(r)} & A_{22}^{(r)} \end{bmatrix}$

and $|\det P_2^{(r)}| = \nu^{(r)} \geq \mu_0^{(r)2} - \mu_1^{(r)2} \geq (1 - \alpha^2) \mu_0^{(r)2}$.

So $A_{ij}^{(r-2)} = \{[(A_2^{(r)})_{ij} - (M_2^{(r)})_{i1} (C_1^{(r)})_{j1} (1 + \epsilon)] (1 + \epsilon) -$

$(M_2^{(r)})_{12} \, (C_2^{(r)})_{j2} \, (1 + \epsilon) \, \}(1 + \epsilon)$. From §10.6, we have

$$G_{1j}^{(r-2)} = \epsilon \ (A_2^{(r)})_{1j} - (M_2^{(r)})_{11} \, (C_2^{(r)})_{j1} \, [(1 + \epsilon)^3 - 1]$$

$$- (M_2^{(r)})_{12} \, (C_2^{(r)})_{j2} \, [(1 + \epsilon)^2 - 1] \ .$$

(10.10.3) $\quad |G_{1j}^{(r-2)}| \le 2^{-t} \, \mu_0^{(r)} \, \{1 + [3 + 0(2^{-t})] \, |(M_2^{(r)})_{11}| +$

$$[2 + 2^{-t}] \, |(M_2^{(r-1)})_{12}| \ \} \ .$$

Now $\quad (M_2^{(r)})_{11} = [\dfrac{A_{11}^{(r)} \, A_{22}^{(r)} \, (1 + \epsilon) - A_{12}^{(r)} \, A_{21}^{(r)} \, (1 + \epsilon)}{A_{22}^{(r)} \, A_{11}^{(r)} \, (1 + \epsilon) - A_{21}^{(r)^2} \, (1 + \epsilon)}] \, (1 + \epsilon) \ .$

Thus $\quad [A_{22}^{(r)} \, A_{11}^{(r)} - A_{21}^{(r)^2}] \, (M_2^{(r)})_{11} = [A_{11}^{(r)} \, A_{22}^{(r)} - A_{12}^{(r)} \, A_{21}^{(r)}] +$

$A_{11}^{(r)} \, A_{22}^{(r)} \, \epsilon \, (2 + \epsilon) - A_{12}^{(r)} \, A_{21}^{(r)} \, \epsilon \, (2 + \epsilon) + A_{21}^{(r)^2} \, \epsilon - A_{22}^{(r)} \, A_{11}^{(r)} \, \epsilon \ .$

From §10.5,

$$|(\Theta_2^{(r)})_{11}| \le [\mu_0^{(r)} \, (\mu_0^{(r)} + \mu_1^{(r)}) \, 2^{-t} \, (2 + 2^{-t}) + \mu_0^{(r)^2} \, 2^{-t} + \mu_1^{(r)^2} \, 2^{-t}]/\nu^{(r)}$$

$$\le \frac{\mu_0^{(r)} \, 2^{-t} \, (2 + 2^{-t})}{\mu_0^{(r)} - \mu_1^{(r)}} + \frac{2^{-t} \, \mu_0^{(r)^2} \, (1+\alpha^2)}{(1-\alpha^2) \, \mu_0^{(r)^2}} \le \{ \frac{2^{-t}}{1-\alpha} \, 2 + 2^{-t} + \frac{1+\alpha^2}{1+\alpha} \}$$

But $\quad (1 + \alpha^2)/(1 + \alpha) < 1 \quad$ since $\quad 0 < \alpha < 1$. Thus we have

(10.10.4) $\quad |(\Theta_2^{(r)})_{11}| \le 2^{-t} \, \{3/(1 - \alpha) + 0(2^{-t})\} \ .$

Similarly, we obtain

(10.10.5) $\quad |(0_2^{(r)})_{12}| \le 2^{-t} \, \{3/(1 - \alpha) + 0(2^{-t})\} \ .$

From §10.6, we recall: $M_2^{(r)} = C_2^{(r)} P_2^{(r)-1} + O_2^{(r)}$. But

(10.10.6) $\left|\left(C_2^{(r)} \left(P_2^{(r)-1}\right)\right)_{ij}\right| \le \mu_0^{(r)} \left(\mu_0^{(r)} + \mu_1^{(r)}\right)/\nu^{(r)} \le 1/(1 - \alpha)$.

Thus from (10.10.4) - (10.10.6), we have for $j = 1,2$:

(10.10.7) $\left|\left(M_2^{(r)}\right)_{ij}\right| \le \{1 + 2^{-t} [3 + 2^{-t}]\}/(1 - \alpha)$.

From (10.10.3) and (10.10.7), we conclude:

(10.10.8) $\left|G_{ij}^{(r-2)}\right| \le 2^{-t} \mu_0^{(r)} \{1 + 5/(1 - \alpha) + O(2^{-t})\}$.

Now we need to bound $\left(M_2^{(r)} P_2^{(r)} O_2^{(r)t}\right)_{ij}$. From §10.6, we recall: $M_2^{(r)} P_2^{(r)} = C_2^{(r)} + O_2^{(r)} P_2^{(r)}$.

From (10.10.4), (10.10.5), and §10.6, we have

(10.10.9) $\left|\left(M_2^{(r)} P_2^{(r)}\right)_{ij}\right| < \mu_0^{(r)} \{1 + 2^{-t} [3(1 + \alpha)/(1 - \alpha) + O(2^{-t})]\}$,

since $\mu_1^{(r)} < \alpha \mu_0^{(r)}$.

From (10.10.4), (10.10.5), and (10.10.9), we conclude:

(10.10.10) $\left|\left(M_2^{(r)} P_2^{(r)} O_2^{(r)t}\right)_{ij}\right| < \{3 + O(2^{-t})\} 2 \mu_0^{(r)} 2^{-t}/(1 - \alpha)$.

10.11   Summary of Floating Point Analysis for F

From §10.7, (10.10.8), and (10.10.10) we have for $i \ge j$:

(10.11.1) $|G_{ij}| 2^t \le [1 + 2/\alpha + O(2^{-t})] \sum_1^' \mu_0^{(k)} +$

$$[1 + 5/(1 - \alpha) + O(2^{-t})] \sum_1^{''} \mu_0^{(k)} \ ,$$

(10.11.2) $\left|\left(M D O^t\right)_{ij}\right| 2^t < [1/\alpha + O(2^{-t})] \sum_1^' \mu_0^{(k)} +$

$$[6/(1 - \alpha) + O(2^{-t})] \sum_1^{''} \mu_0^{(k)} \ .$$

From (10.11.1) and (10.11.2), we conclude for $i \geq j$:

(10.11.3) $\quad |F_{ij}| \; 2^t < [1 + 3/\alpha + O(2^{-t})] \; \sum_1^{'} \mu_0^{(k)} +$

$$[1 + 11/(1 - \alpha) + O(2^{-t})] \; \sum_1^{''} \mu_0^{(k)} \quad .$$

We shall now assume that the first $O(2^{-t})$ term in (10.11.3) is bounded by 0.01 (which is true for $t \geq 8$ for all α) and that the second $O(2^{-t})$ term in (10.11.3) is bounded by 0.02 (which is true for $t \geq 14$ when $\alpha = \alpha_0 = (1 + \sqrt{17}/8))$.

Under these assumptions, we have:

(10.11.4) $\quad |F_{ij}| \; 2^t < [1 + (3.01)/\alpha] \; \sum_1^{'} \mu_0^{(k)} +$

$$[1 + (11.02)/(1 - \alpha)] \; \sum_1^{''} \mu_0^{(k)} \quad ,$$

while for $\alpha = \alpha_0$, we have:

(10.11.5) $\quad |F_{ij}| \; 2^t < 5.71 \; \sum_1^{'} \mu_0^{(k)} + 3.16 \; \sum_1^{''} \mu_0^{(k)} \quad .$

10.12   Comments on the Bound for F

From (10.11.4), (10.11.5), and §10.3, we see that we have a good bound on F (and hence on E) only if the maximal elements $\mu_0^{(k)}$ in the reduced matrices $\mathtt{A}^{(k)}$ do not grow too large.

From Corollary 3 of §5.7, we have for $\alpha = \alpha_0$:

$\mu_0^{(k)} < (2.57)^{n-k} \mu_0$. Thus (10.11.5) does not automatically give a good bound on F if $n \geq t$. At the present time, most computers have $t \doteq 50$. Hence (10.11.5) cannot guarantee a good bound on F for systems of order ≥ 50 .

According to Wilkinson (1965, p. 215), if $LU = A + F$, then

$$\max_{i,j} |F_{ij}| \le (2.01)\, n\, 2^{-t} \max_k \mu_0^{(k)} \quad \text{for Gaussian elimination, where}$$

$\max_k \mu_0^{(k)} \le 2^n\, \mu_0$ for partial pivoting, while $\max_k \mu_0^{(k)} \le \sqrt{n}\, f(n)\, \mu_0$

for complete pivoting.

In Chapter 11, we shall show that for the diagonal pivoting method with $0 < \alpha < 1$, we have:

$$\max_k \mu_0^{(k)} \le \sqrt{n}\, f(n)\, \mu_0\, c(\alpha)\, h(n,\alpha) \ .$$

In Chapter 12, we shall show that for $\alpha = \alpha_0$:

$$c(\alpha_0)\, h(n,\alpha_0) < 3.07\, (n-1)^{0.446} \quad \text{for} \quad n \ge 2 \ .$$

Then from (10.11.5) we have for $\alpha = \alpha_0$:

$$\max_{i,j} |F_{ij}| \le 2^{-t}\, \sqrt{n}\, f(n)\, \mu_0\, (3.07)n^{0.446}\, [5.71p + 31.6(n-p)/2] \ ,$$

where p is the number of 1×1 pivots used. So

$$\max_{i,j} |F_{ij}| \le 2^{-t}\, \sqrt{n}\, f(n)\, \mu_0\, (3.07)n^{0.446}\, [15.8n - 10.09p]$$

$$\le 15.8n\, 2^{-t}\, \sqrt{n}\, f(n)\, \mu_0\, (3.07)n^{0.446} \ ,$$

which is within a factor $7.9(3.07)n^{0.446}$ of the bound for $\max_{i,j} |F_{ij}|$

for Gaussian elimination with complete pivoting.

10.13 Floating Point Analysis for $\partial\, D$

The blocks of D are of order 1 or 2 .

(a) Suppose $[D_{ii}]$ is a block of order 1. Then so is $[(\delta\, D)_{ii}]$.

We want to solve $D_{ii}\, y_i = c_i$, but we obtain $(D_{ii} + (\delta\, D)_{ii})y_i = c_i$

due to the finite precision; $\text{pivot}[n-i+1] = 1$ and $|D_{ii}| = \mu_1^{(n-i+1)}$.

Thus $y_i = (c_i/D_{ii})(1 + \epsilon)$ and $(\delta D)_{ii} = - D_{ii} \epsilon$.

So $|(\delta D)_{ii}| \leq 2^{-t} |D_{ii}|$. But $|D_{ii}| = \mu_1^{(n-i+1)} \leq \mu_0^{(n-i+1)}$.

Thus $|(\delta D)_{ii}| \leq 2^{-t} \mu_0^{(n-i+1)}$.

(b) Suppose $\begin{bmatrix} D_{ii} & D_{i+1,i} \\ D_{i+1,i} & D_{i+1,i+1} \end{bmatrix} = P_2^{(n-i+1)}$ is a block of order

2. Then pivot$[n-i+1] = 2$ and $\mu_1^{(n-i+1)} < \alpha \, \mu_0^{(n-i+1)}$.

$\pm \, \nu^{(n-k+1)} = [D_{ii} \, D_{i+1,i+1} \, (1 + \eta) - D_{i+1,i}^2 \, (1 + \eta)] \, (1 + \epsilon)$,

where $\eta = 2^{-t} \, 0(2^{-t})$. Since $|\epsilon| \leq 2^{-t}$, we have

$|\nu^{(n-k+1)} - \det P_2^{(n-i+1)}| \leq 2^{-t} \, [1 + 0(2^{-t})] \, (1 + \alpha^2) \, \mu_0^{(n-i+1)^2}$.

Now $y_{i+1} = [c_{i+1} \, D_{ii} \, (1 + \eta) - c_i \, D_{i+1,i} \, (1 + \eta)](1 + \epsilon)/(\pm \nu^{(n-k+1)})$

and $y_i = [c_i \, D_{i+1,i+1}(1+\eta) - c_{i+1} \, D_{i+1,i} \, (1+\eta)] \, (1+\epsilon)/(\pm \nu^{(n-k+1)})$.

Since $|\epsilon| \leq 2^{-t}$, $|\eta| \leq 2^{-t} \, 0(2^{-t})$, and $\mu_1^{(n-i+1)} < \alpha \, \mu_0^{(n-i+1)}$,

after much manipulation, we have:

$$\begin{bmatrix} |(\delta D)_{ii}| & |(\delta D)_{i,i+1}| \\ |(\delta D)_{i+1,i}| & |(\delta D)_{i+1,i+1}| \end{bmatrix} \leq 2^{-t} \, [1 + 0(2^{-t})] \, \mu_0^{(n-i+1)} \begin{bmatrix} \alpha & 1 \\ 1 & \alpha \end{bmatrix} .$$

10.14 Floating Point for M_1 and M_2

From §10.2, $(M + M_1)c = b$ and $(M + M_2)^t x = y$.

From Wilkinson (1965, pp. 247-249), we have:

$$|(M_1)_{ii}| \;,\; |(M_2)_{ii}| \leq 2^{-t} [1 + O(2^{-t})] |M_{ii}| \;,\; (M_1)_{ij} = 0 = (M_2^t)_{ij}$$

for $i < j$,

$$|(M_1)_{ij}| \;,\; |(M_2^t)_{ij}| \leq \frac{3}{2} 2^{-t} [1 + O(2^{-t})] (n - 2 + i - j) |M_{ij}|$$

for $i > j$.

From §10.10, $M_{ii} = 1$ and for $i > j$:

$$|M_{ij}| \leq \max \{1/\alpha, 1/(1-\alpha)\} [1 + O(2^{-t})] .$$

So for $i > j$:

$$(10.14.1) \quad |(M_1)_{ij}|, \; |(M_2^t)_{ij}| \leq 2^{-t} \max\{1/\alpha, 1/(1-\alpha)\}[1+O(2^{-t})]3(n-2+i-j)/2$$

and $\quad |(M_1)_{ii}| \;,\; |(M_2)_{ii}| \leq 2^{-t} [1 + O(2^{-t})]$.

10.15 Floating Point Bound for E

From §§10.3, 10.11, 10.13, and 10.14, we could express E_{ij} in terms of F_{ij} and the elements of M, D, M_1, M_2, and δD . But this expression is very complicated.

Let us define $\|A\| = \max_j \sum_{i=1}^{n} |A_{ij}|$ where A is $n \times n$. (This is usually called the one-norm.) Clearly, $|A_{ij}| \leq \|A\|$.

From §10.3, we now have

$$(10.15.1) \quad |E_{ij}| \leq |F_{ij}| + \varepsilon \;, \text{ where}$$

$$\varepsilon = \|M_1\| \; \|D + \delta D\| \; \|(M+M_2^t)\| + \|M\| \; [\|\delta D\| \; \|(M+M_2^t)\| + \|D\| \; \|M_2^t\|] .$$

$\|D\| \leq (1 + \alpha) \max_k \mu_0^{(k)}$. From §10.13,

$\|\delta D\| \leq 2^{-t} (1 + \alpha) \max_k \mu_0^{(k)} [1 + 0(2^{-t})]$.

From §10.10, $\|M\|$, $\|M^t\| \leq 1 + (n-1) \max \{1/\alpha, 1/(1-\alpha)\} [1 + 0(2^{-t})]$.

Let us assume $0(2^{-t}) (1 + \max \{1/\alpha, 1(1-\alpha)\} 3(n + 2)^2) = 0(2^{-t})$,

i.e. let us assume $n^2 2^{-t} << 1$. Then

$\|M_1\|$, $\|M_2^t\| \leq 2^{-t} [1 + 0(2^{-t})]$.

Now we can bound ϵ . Let $\beta = 1 + (n-1) \max \{1/\alpha, 1/(1-\alpha)\}$.

(10.15.2) $\epsilon \leq 2^{-t} (1 + \alpha) \max_k \mu_0^{(k)} \beta(2 + \beta) [1 + 0(2^{-t})]$.

For $\alpha = \alpha_0$, we have $\beta < 2.781n$, so

(10.15.3) $\epsilon \leq 2^{-t} \max_k \mu_0^{(k)} (7.734)n^2 [1 + 0(2^{-t})]$.

From (10.11.4), (10.15.1), and (10.15.2) we conclude:

(10.15.4) $|E_{ij}| 2^t < [1 + 3.0/\alpha] \sum_i' \mu_0^{(k)} + [1 + 11.02/(1-\alpha)] \sum_i'' \mu_0^{(k)}$

$+ (1 + \alpha) \max_k \mu_0^{(k)} \beta(2 + \beta)[1 + 0(2^{-t})]$.

From (10.11.5), (10.15.1), and (10.15.3) we conclude for

$\alpha = \alpha_0$:

(10.15.5) $|E_{ij}| < 2^{-t} \max_k \mu_0^{(k)} \{15.80n + 7.74 n^2\}$.

Since $\|E\| \leq \|F\| + \epsilon$, from (10.11.5) and (10.15.3):

(10.15.6) $\|E\| < (23.54)n^2 2^{-t} \max_k \mu_0^{(k)}$.

10.16   Comments on the Bound for E

From §10.12 and §10.15, we need to bound the $\mu_0^{(k)}$ in order to
have a good bound or E .

In Chapter 11 we shall show that $\max_k \mu_0^{(k)} \leq \sqrt{n}\ f(n)\ \mu_0\ c(\alpha)\ h(n,\alpha)$

for the diagonal pivoting method for $0 < \alpha < 1$. In Chapter 12 we
shall show that $c(\alpha_0)\ h(n,\alpha_0) < 3.07(n-1)^{0.446}$. Then we would have

$$\|E\| < (23.54)n^2\ 2^{-t}\ \sqrt{n}\ f(n)\ \mu_0\ (3.07)(n-1)^{0.446} .$$

For Gaussian elimination, in order to solve $A x = b$, in finite
precision we actually perform:

(1) $L U = A + F$

(2) $c = (L + \delta L)^{-1}\ b$

(3) $x = (U + \delta U)^{-1}\ c.$

From Wilkinson (1965, p. 215, pp. 248-252), we have (cf. §2.5):

$$\max_{i,j}\ |F_{ij}| \leq 2.01n\ 2^{-t}\ \max_k \mu_0^{(k)} \quad \text{and}$$

$$\|E\| < 2.01n^2\ 2^{-t}\ \max_k\ \mu_0^{(k)}\ [1 + 0(2^{-t})] .$$

For complete pivoting, $\max_k \mu_0^{(k)} \leq \sqrt{n}\ f(n)\ \mu_0$ (Wilkinson, 1961).

Thus our bound on $\|E\|$ for diagonal pivoting differs by a factor

$36(n-1)^{0.446}$ from the bound on $\|E\|$ for Gaussian elimination with
complete pivoting.

Chapter 11 : An A Posteriori Bound on Element

Growth for $0 < \alpha < 1$

11.1 Introduction

We saw in Chapter 10 that the bound on the elements of the error
matrix depends on the maximum of the elements of the reduced matrices.

Let $A = A^{(n)}$ be the original symmetric matrix of order n with
det $A \neq 0$. Let $A^{(k)}$ be the reduced matrix of order k .

We saw in §8.4 that $\max\limits_{i,j} |A^{(k)}_{ij}| < (2.57)^{n-k} \max\limits_{i,j} |A_{ij}|$. But we
shall show in §§11.2–11.7 that we can get a much better bound by the
use of Wilkinson's techniques for Gaussian elimination with complete
pivoting (Wilkinson, 1961; pp. 281-285).

His proof depended on the fact that the pivots in Gaussian eli-
mination with complete pivoting were maximal elements in the reduced
matrices. Our pivots are not necessarily maximal elements, but they
are closely related to the maximal elements in the reduced matrices.

We shall assume that we use strategy S_α (§5.7) for any α ,
$0 < \alpha < 1$, and that $v^{(k)} \geq \mu_0^{(k)2} - \mu_1^{(k)2}$ for all $A^{(k)}$ (see
Lemma 4, §5.5). In particular, this lower bound holds for the strate-
gies in §6.1, §6.4, and §8.1.

11.2 The Pivots

Let $A^{(k)}$ be the reduced matrix of order k .

Let $\mu_0^{(k)} = \max_{i,j} |A_{ij}^{(k)}|$ and $\mu_1^{(k)} = \max_{i} |A_{ii}^{(k)}|$.

Let us assume that, whenever we shall use a 2×2, interchanges have ensured that $\nu^{(k)} = |A_{11}^{(k)} A_{22}^{(k)} - A_{21}^{(k)^2}| \geq \mu_0^{(k)^2} - \mu_1^{(k)^2}$.

(In particular, this assumption holds for the strategies in §§6.1-6.2, §§6.4-6.5, and §§8.1-8.2.)

From strategy S_α (§5.7) and §9.1, we recall:

$$\text{pivot}[k] = \begin{cases} 1 & \text{if } \mu_1^{(k)} \leq \alpha \mu_0^{(k)} \text{ , i.e. if } A^{(k)} \text{ uses a 1×1 pivot} \\ 2 & \text{if } \mu_1^{(k)} < \alpha \mu_0^{(k)} \text{ , i.e. if } A^{(k)} \text{ uses a 2×2 pivot} \\ 0 & \text{if } \mu_1^{(k+1)} < \alpha \mu_0^{(k+1)} \text{ , i.e. if } A^{(k+1)} \text{ uses a 2×2 pivot.} \end{cases}$$

Let us now define

$$p_k = \begin{cases} \mu_1^{(k)} & \text{if } \text{pivot}[k] = 1 \\ \sqrt{\nu^{(k)}} & \text{if } \text{pivot}[k] = 2 \\ \sqrt{\nu^{(k+1)}} & \text{if } \text{pivot}[k] = 0 \end{cases}$$

The p_k will be called pivots.

From the decomposition $A = M D M^t$ we see that $|\det A| = p_1 \cdots p_n$ and $|\det A^{(k)}| = p_1 \cdots p_{k'}$, since $\det M = 1$.

11.3   Hadamard's Inequality

By Hadamard's Inequality (Gantmacher, pp. 253-254),

$$(11.3.1) \quad |\det A| \leq \{ \prod_{i=1}^{n} \sum_{j=1}^{n} A_{ij}^2 \}^{1/2} \leq (n \mu_0^{(n)^2})^{n/2} = (\sqrt{n} \mu_0^{(n)})^n \text{ ,}$$

since $\mu_0^{(n)} = \mu_0 = \max_{i,j} |A_{ij}|$. Also

$$(11.3.2) \quad |\det A^{(k)}| \leq (\sqrt{k} \mu_0^{(k)})^k \text{ .}$$

11.4   Bounding $\det A^{(k)}$

(1) If $\text{pivot}[k] = 1$: $\mu_1^{(k)} \geq \alpha\,\mu_0^{(k)}$. So $\mu_0^{(k)} \leq p_k/\alpha$.

(11.4.1) $p_1 \cdots p_k = |\det A^{(k)}| \leq (\sqrt{k}\,\mu_0^{(k)})^k \leq (p_k\,\sqrt{k}/\alpha)^k$.

(2) If $\text{pivot}[k] = 2$: $\mu_1^{(k)} < \alpha\,\mu_0^{(k)}$. If we assume

$\nu^{(k)} \geq \mu_0^{(k)2} - \mu_1^{(k)}$ for all $A^{(k)}$, then $\nu^{(k)} \geq (1 - \alpha^2)\,\mu_0^{(k)2}$.

Thus $\mu_0^{(k)2} \leq \nu^{(k)}/(1 - \alpha^2) = p_k^2/(1 - \alpha^2)$.

(11.4.2) $p_1 \cdots p_k = |\det A^{(k)}| \leq (p_k\,\sqrt{k/(1 - \alpha^2)}\,)^k$.

11.5   Fundamental Inequality

From §11.4 (and later in §11.5), we are led to the following:

Definition:

$$(11.5.1)\quad \beta_k = \begin{cases} 1/\alpha^2 & \text{if } \text{pivot}[k] = 1 \\[2mm] 1/(1-\alpha^2) & \text{if } \text{pivot}[k] = 2 \ . \\[2mm] \dfrac{1}{k}\,[\dfrac{k+1}{1-\alpha^2}]^{1+1/k} & \text{if } \text{pivot}[k] = 0 \end{cases}$$

From (11.4.1) and (11.4.2) we have:

(11.5.2) $p_1 \cdots p_k \leq (\sqrt{k\,\beta_k}\,p_k)^k$ if $\text{pivot}[k] \neq 0$, $1 \leq k \leq n$.

We would like to have a similar equation for $\text{pivot}[k] = 0$ for our analysis in §11.6.

If $\text{pivot}[k] = 0$, then $\text{pivot}[k+1] = 2$. By (11.5.2),

$p_1 \cdots p_k\, p_{k+1} \leq (\sqrt{(k+1)\,\beta_{k+1}}\,p_{k+1})^{k+1}$, where $\beta_{k+1} = 1/(1 - \alpha^2)$ and

$p_k = p_{k+1}$. Since $\beta_k = \dfrac{1}{k}\,[\dfrac{k+1}{1-\alpha^2}]^{1+1/k}$ by (11.5.1),

$p_1 \cdots p_k \le [(k+1) \beta_{k+1}]^{(k+1)/2} p_k^k = (\sqrt{k \beta_k} \, p_k)^k$. Thus

(11.5.3) $\quad p_1 \cdots p_k \le (\sqrt{k \beta_k} \, p_k)^k$ for all k , $1 \le k \le n$.

11.6   Bounding Pivot Growth

Define $q_k = \log p_k$. From (11.5.3),

(11.6.1) $\quad \displaystyle\sum_{i=1}^{k-1} q_i \le (k-1) q_k + \frac{1}{2} k \log (k \beta_k)$.

Since $\left| \det A^{(n)} \right| = p_1 \cdots p_n$, we have:

(11.6.2) $\quad \log \left| \det A^{(n)} \right| = \displaystyle\sum_{i=1}^{n} q_i$.

Dividing (11.6.1) by $k(k-1)$ for $2 \le k \le n-1$ and (11.6.2) by $n-1$ and adding we have:

$q_1 + q_2/2 + q_3/3 + \ldots q_{n-2}/(n-2) + q_{n-1}/(n-1) + q_n/(n-1)$

$\le \dfrac{1}{n-1} \log \left| \det A^{(n)} \right| + \dfrac{1}{2} \displaystyle\sum_{k=2}^{n-1} \log (k \beta_k)^{1/(k-1)}$

$\quad\quad + q_2/2 + q_3/3 + \ldots + q_{n-1}/(n-1)$,

after observing that $\displaystyle\sum_{r=k}^{n-1} \frac{1}{r(r-1)} = \frac{1}{k-1} - \frac{1}{n-1}$.

(11.6.3) $\quad q_1 + q_n/(n-1) \le \dfrac{1}{2} \log \displaystyle\prod_{k=2}^{n-1} (k \beta_k)^{1/(k-1)} + \dfrac{1}{n-1} \log \left| \det A^{(n)} \right|$.

From (11.5.3), $\left| \det A^{(n)} \right| \le (\sqrt{n \beta_n} \, p_n)^n$. We define

(11.6.4) $\quad f(r) = \{ \displaystyle\prod_{k=2}^{r} k^{1/(k-1)} \}^{1/2}$,

(11.6.5) $\quad h(r,\alpha) = \{ \displaystyle\prod_{k=2}^{r} \beta_k^{1/(k-1)} \}^{1/2}$.

From (11.6.3) we now have

$$q_1 + q_n/(n-1) \leq \log f(n-1) + \log h(n-1,\alpha) + \frac{n}{2(n-1)} \log (n \beta_n) + n q_n/(n-1) \ .$$

With simple manipulation we have

$$q_1 - q_n \leq \log f(n-1) + \frac{1}{2(n-1)} \log n + \log h(n-1,\alpha)$$

$$+ \frac{1}{2(n-1)} \log \beta_n + \frac{1}{2} \log (n \beta_n)$$

$$= \log f(n) + \log h(n,\alpha) + \frac{1}{2} \log (n \beta_n) \ .$$

Thus we conclude

(11.6.6) $p_1/p_n \leq \sqrt{n} \ f(n) \ \sqrt{\beta_n} \ h(n,\alpha) \ .$

Similarly, we have for $1 \leq k \leq n$:

(11.6.7) $p_k/p_n \leq \sqrt{n-k+1} \ f(n-k+1) \ \sqrt{\beta_{n-k+1}} \ h(n-k+1,\alpha)$

$$\leq \sqrt{n} \ f(n) \ \sqrt{\beta_{n-k+1}} \ h(n,\alpha) \ .$$

(11.6.6) and (11.6.7) hold for all α , $0 < \alpha < 1$, under strategy S_α (§5.7) provided that $\nu^{(k)} \geq \mu_0^{(k)^2} - \mu_1^{(k)^2}$ for all $A^{(k)}$ (§5.5).

We now have a bound on pivot growth. But we are interested in bounding element growth (§11.1). If A finishes with a 2×2 pivot we are interested in bounding $\mu_0^{(2)}$, while if A finishes with a 1×1 pivot, we must bound $\mu_0^{(1)}$.

11.7   Bounding Element Growth

We shall now express p_1, p_2, $\sqrt{\beta_n} \ p_n$, and $\sqrt{\beta_{n-1}} \ p_n$ in terms of $\mu_0^{(1)}$, $\mu_0^{(2)}$, and $\mu_0^{(n)}$.

Let $A = A^{(k)}$, $\mu_0 = \mu_0^{(n)}$, $\mu_1 = \mu_1^{(n)}$, $\nu = \nu^{(n)}$.

Since pivot$[n] \neq 0$, $P_n = \begin{cases} \mu_1 & \text{if pivot}[n] = 1 \\ \\ \sqrt{\nu} & \text{if pivot}[n] = 2 \end{cases}$, and

$$\beta n = \begin{cases} 1/\alpha^2 & \text{if pivot}[n] = 1 \\ \\ 1/(1 - \alpha^2) & \text{if pivot}[n] = 2 \end{cases}.$$

But $\mu_1 \leq \mu_0$ always. If pivot$[n] = 2$, then $\mu_1 < \alpha \mu_0$, and

(by §5.5) $\sqrt{\nu} \leq \sqrt{\mu_0^2 + \mu_1^2} \leq \sqrt{1 + \alpha^2}\, \mu_0$. Thus

$$P_n \leq \begin{cases} \mu_0 & \text{if pivot}[n] = 1 \\ \\ \sqrt{1 + \alpha^2}\, \mu_0 & \text{if pivot}[n] = 2 \end{cases}.$$

Now $\beta_{n-1} = \begin{cases} 1/\alpha^2 & \text{if pivot}[n-1] = 1 \\ 1/(1 - \alpha^2) & \text{if pivot}[n-1] = 2 \\ \dfrac{n\,\beta n}{n-1}\, (n\,\beta n)^{1/(n-1)} & \text{if pivot}[n-1] = 0 \end{cases}$.

Let us define:

(11.7.1) $m(\alpha) = \max \{\frac{1}{\alpha}, \sqrt{(1+\alpha^2)/(1-\alpha^2)}\,\}$.

Thus we have:

(11.7.2) $\sqrt{\beta n}\, P_n \leq m(\alpha)\, \mu_0$, and

(11.7.3) $\sqrt{\beta_{n-1}}\, P_n \leq \begin{cases} m(\alpha)\, \mu_0 & \text{if pivot}[n] = 1 \\ \\ (\sqrt{n\,\beta_n})^{1/(n-1)}\sqrt{n/(n-1)}\ \sqrt{(1+\alpha^2)/(1-\alpha^2)}\, \mu_0 & \text{if pivot}[n] = 2 \end{cases}$

From (11.6.4) and (11.6.5), we have

(11.7.4) $(\sqrt{n\,\beta_n})^{1/(n-1)}\, f(n-1)\, h(n-1,\alpha) = f(n)\, h(n-1,\alpha)$.

We have two cases (§11.6) as to whether:

(a) the last pivot is 1×1, i.e. pivot[1] = 1 ,

(b) the last pivot is 2×2, i.e. pivot[2] = 2 .

 Case (a) : Let pivot[1] = 1 . Then $p_1 = \mu_1^{(1)} = \mu_0^{(1)}$.

From (11.6.6) and (11.7.2) we have

(11.7.5) $\mu_0^{(1)} = p_1 \leq \sqrt{n} \; f(n) \; \mu_0 \; m(\alpha) \; h(n,\alpha)$.

 Case (b) : Let pivot[2] = 2 . Then $p_2 = \sqrt{\nu^{(2)}}$ and

$\mu_1^{(2)} < \alpha \; \mu_0^{(2)}$. Since we have assumed $\nu^{(2)} \geq \mu_0^{(2)^2} - \mu_1^{(2)^2}$,

$\nu^{(2)} \geq (1 - \alpha^2) \; \mu_0^{(2)^2}$. Hence

(11.7.6) $\mu_0^{(2)} \leq p_2/\sqrt{1 - \alpha^2}$.

From (11.6.7) with k = 2 and from (11.7.6) we have

(11.7.7) $\mu_0^{(2)} \leq \sqrt{n-1} \; f(n-1) \; h(n-1,\alpha) \; \sqrt{\beta_{n-1}} \; p_n/\sqrt{1 - \alpha^2}$.

Now we have two cases as to whether the first pivot is (i) 1×1 ,
or (ii) 2×2.

(i) Let pivot[n] = 1 . From (11.7.3) and (11.7.),

$\mu_0^{(2)} \leq \sqrt{n-1} \; f(n-1) \; h(n-1,\alpha) \; m(\alpha) \; \mu_0/\sqrt{1 - \alpha^2}$.

But $f(n-1) < f(n)$ and $h(n-1,\alpha) < h(n,\alpha)$. Thus

(11.7.8) $\mu_0^{(2)} \leq \sqrt{n} \; f(n) \; \mu_0 \; m(\alpha) \; h(n,\alpha)/\sqrt{1 - \alpha^2}$.

(ii) Let pivot[n] = 2 . From (11.7.3), (11.7.4), and (11.7.7),

$\mu_0^{(2)} \leq \sqrt{n} \; f(n) \; \mu_0 \; h(n,\alpha) \; \sqrt{1 + \alpha^2}/(1 - \alpha^2)$.

But $\sqrt{1 + \alpha^2}/(1 - \alpha^2) \leq m(\alpha)/\sqrt{1 - \alpha^2}$ from (11.7.1). Hence

$$(11.7.9) \quad \mu_0^{(2)} \leq \sqrt{n} \ f(n) \ \mu_0 \ m(\alpha) \ h(n,\alpha)/\sqrt{1 - \alpha^2} \ .$$

Let us now define:

$$(11.7.10) \quad c(\alpha) = m(\alpha) \times \begin{cases} 1 & \text{if pivot}[1] = 1 \\ 1/\sqrt{1 - \alpha^2} & \text{if pivot}[2] = 2 \end{cases} ,$$

where $m(\alpha) = \max \{ 1/\alpha, \sqrt{(1+\alpha^2)/(1-\alpha^2)} \} $.

Then we conclude for $0 < \alpha < 1$:

$$(11.7.11) \quad \begin{cases} \text{If pivot}[1] = 1 : \mu_0^{(1)} \\ \text{If pivot}[2] = 2 : \mu_0^{(2)} \end{cases} \leq \sqrt{n} \ f(n) \ \mu_0 \ c(\alpha) \ h(n,\alpha) \ .$$

Similarly, we have for each reduced matrix $A^{(k)}$,

$$(11.7.12) \quad \mu_0^{(k)} \leq \sqrt{n - k+j} \ f(n - k+j) \ \mu_0 \ c(k,\alpha) \ h(n - k+j, \ \alpha) \ ,$$

where $j = \text{pivot}[k]$, $c(k,\alpha) = m(\alpha) \times \begin{cases} 1 & \text{if } j = 1 \\ 1/\sqrt{1 - \alpha^2} & \text{if } j = 2 \end{cases} ,$

and $0 < \alpha < 1$.

Hence, for all $A^{(k)}$, we have for $0 < \alpha < 1$:

$$(11.7.13) \quad \mu_0^{(k)} \leq \sqrt{n} \ f(n) \ \mu_0 \ c(\alpha) \ h(n,\alpha) \ .$$

11.8   Comments on the Bound

The bound in (11.7.13) holds under strategy S_α , $0 < \alpha < 1$, and under the assumption $v^{(k)} \geq \mu_0^{(k)^2} - \mu_1^{(k)^2}$ for all $A^{(k)}$. In particular, (11.7.13) holds for the unequilibrated diagonal pivoting strategy (§§8.1-8.2), the complete strategy (§§6.1-6.3), and the partial

equilibrated strategy (§§6.4–6.5). (For the partial equilibrated stra-
tegy, we assume that, given a reduced matrix $A^{(k)}$ with $\mu_0^{(k)} = \max_{i,j} |A_{ij}^{(k)}|$,

we equilibrated $A^{(k)}$ so that the maximal element in absolute value in

each row is $\mu_0^{(k)}$.)

Wilkinson (1961) obtains $\sqrt{n}\, f(n)\, \mu_0$ as the bound on the elements

in the reduced matrices for solving $A\,x = b$, $\mu_0 = \max_{i,j} |A_{ij}|$, by

Gaussian elimination with complete pivoting. We have the extra factor

$c(\alpha)\, h(n,\alpha)$ since our pivots are not necessarily maximal elements of

the reduced matrices.

We call the bounds in (11.7.11) – (11.7.13) a posteriori, since

we cannot calculate the β_k terms of $h(n,\alpha)$ until we know the posi-

tion of the blocks of order 1 and 2 in D for the decomposition

$A = M\,D\,M^t$. In Chapter 12, we shall give a bound on $c(\alpha)\, h(n,\alpha)$,

independent of the structure of D , for the value $\alpha = \alpha_0 =$

$(1 + \sqrt{17})/8$ (§5.7).

11.9   Smaller Bound on Pivot Growth

If H is an $n \times n$ positive-definite Hermitian matrix and λ_{min}

is the minimum eigenvalue of A , then (Shisha, p. 173):

$$(11.8.1) \quad \det H \leq \prod_{i=1}^{n} H_{ii} - \lambda_{min}^{n-2} \Big(\sum_{n \geq j > i \geq 1} |H_{ij}|^2 \Big) .$$

If A is an $n \times n$ non-singular real symmetric matrix and λ is

the minimum of the absolute values of the eigenvalues of A , then,

setting $H = A^t A$ in the above, we have

(11.8.2) $\quad |\det A|^2 \le \prod_{i=1}^{n} \left(\sum_{j=1}^{n} A_{ij}^2 \right) - \lambda^{2(n-2)} \sum_{n \ge j > i \ge 1} |A_i^t A_j|^2$,

where A_j is the jth column of A .

Then $\quad |\det A|^2 \le (n \mu_0^2) \epsilon(A)$, where

(11.8.3) $\quad \epsilon(A) = 1 - \lambda^{2(n-2)} \left[\sum_{n \ge j > i \ge 1} |A_i^t A_j|^2 \right] / (n \mu_0^2)^n$.

Using an analysis similar to that in §§11.2-11.7, we obtain:

(11.8.4) $\quad \left\{ \begin{array}{l} \text{If pivot[1]} = 1 : \mu_0^{(1)} \\ \\ \text{If pivot[2]} = 2 : \mu_0^{(2)} \end{array} \right\} \le \sqrt{n} \; f(n) \; \mu_0 \; c(\alpha) \; h(n,\alpha) \; \tau(A)$,

where $\quad \tau(A)^2 = \epsilon(A)^{\frac{1}{n-1}} \prod_{r=2}^{n} \epsilon(A^{(r)})^{\frac{1}{r(r-1)}} \quad$ and

$\epsilon(A^{(r)}) = 1 - \lambda_r^{2(r-2)} \left[\sum |A_i^{(r)t} A_j^{(r)}|^2 \right] / [r \mu_0^{(r)2}]^r \quad$, where $A^{(r)}$

is the reduced matrix of order r $(A = A^{(n)})$, λ_r is the minimum of

the absolute values of the eigenvalues of $A^{(r)}$, and $A_j^{(r)}$ is the

j^{th} column of $A^{(r)}$.

If $\det A \ne 0$, then $|\lambda_r| > 0$ for each $A^{(r)}$. If $A^{(r)}$ is

not Hadamard (see §12.6), then $\epsilon(A^{(r)}) < 1$. If $A^{(r)}$ is Hadamard,

then $A^{(r)}$ will use a 1×1 pivot and $A^{(r-1)}$ will not be Hadamard

(see Appendix A).

Thus $\tau(A) < 1$ if $\det A \ne 0$. Hence (11.8.4) gives a lower bound

than does (11.7.11), but (11.8.4) is so complicated we are unable to use

it to advantage, and we present it merely for its academic interest.

Similarly, for Gaussian elimination with complete pivoting, we can obtain $\sqrt{n}\ f(n)\ \mu_0\ \tau(A)$ as the bound on the elements in all the reduced matrices when solving $A\ x = b$, $\mu_0 = \max_{i,j} |A_{ij}|$, $\det A \neq 0$, if we replace the λ_r in $\tau(A)$ by $|\sigma_r|$, the singular value σ_r of $\Lambda^{(r)}$ of minimum modulus.

Chapter 12 : <u>An A Priori Bound on Element Growth</u>

<u>for</u> $\alpha = \alpha_0 = (1 + \sqrt{17})/8$

12.1 <u>Introduction</u>

In (11.7.13) we obtained $\sqrt{n}\ f(n)\ \mu_0\ c(\alpha)\ h(n,\alpha)$ as a bound on the element growth in the reduced matrices, for any α with $0 < \alpha < 1$.

From (11.5.1), (11.6.4), (11.6.5), (11.7.1), and (11.7.10) we recall:

$$\beta_k = \begin{cases} 1/\alpha^2 & \text{if pivot}[k] = 1 \\[2mm] 1/(1 - \alpha^2) & \text{if pivot}[k] = 2 \\[2mm] \frac{1}{k}\ (\frac{k+1}{1-\alpha^2})^{1+1/k} & \text{if pivot}[k] = 0 \end{cases}$$

$$f(n)^2 = \prod_{k=2}^{n} k^{1/(k-1)} \quad,\quad h(n,\alpha)^2 = \prod_{k=2}^{n} \beta_k^{1/(k-1)} \quad,$$

$$m(\alpha) = \max\{\frac{1}{\alpha^2}, \frac{1}{1-\alpha^2}\}, \quad\text{and}\quad c(\alpha) = m(\alpha) \times \begin{cases} 1 & \text{if pivot}[1] = 1 \\[2mm] 1/\sqrt{1 - \alpha^2} & \text{if pivot } [2] = 2 \end{cases}$$

The term $c(\alpha)\ h(n,\alpha)$ arose from the fact that our pivots are not necessarily the maximal elements of the reduced matrices, but can be expressed as multiples (involving α) of such maximal elements. The bound in (11.7.13) is a posteriori since we cannot calculate $c(\alpha)\ h(n,\alpha)$ until we know the pivot selection, i.e. until we know the position of the blocks of order 1 and 2 in the block diagonal matrix D for our decomposition $A = M\ D\ M^t$ for a given value of α, $0 < \alpha < 1$.

We would like an a priori bound on the element growth, i.e. a bound independent of the selection of a 1×1 or a 2×2 pivot at each stage.

We would also like the a priori bound to hold for all α, $0 < \alpha < 1$. But no such bound exists, as we shall now show.

Let p be the number of 1×1 pivots for the decomposition (§9.1), i.e. there are p blocks of order 1 in D above. If $p = n$ (e.g. for a positive definite matrix, see Appendix A), then

$$h(n,\alpha) = \prod_{k=2}^{n} (1/\alpha)^{1/(k-1)} \to \infty \text{ as } \alpha \to 0. \text{ If } p = 0 \text{ (see §6.3), then}$$

$$h(n,\alpha) > \prod_{k=2}^{n} (1/\sqrt{1 - \alpha^2})^{1/(k-1)} \to \infty \text{ as } \alpha \to 1.$$

Thus we shall give an a priori bound for element growth only for the value $\alpha = \alpha_0 = (1 + \sqrt{17})/8$ (§5.7).

12.2 Lower Bound on $c(\alpha)\, h(n,\alpha)$ for $0 < \alpha < 1$

We shall now find a lower bound on $c(\alpha)\, h(n,\alpha)$ for all α, $0 < \alpha < 1$, in order to show that the upper bound for $\alpha = \alpha_0$ that we obtain in §12.4 is a reasonable bound, i.e. that some other choice of α would not provide us with a much better upper bound on $c(\alpha)\, h(n,\alpha)$.

Since $\min_{0<\alpha<1} c(\alpha)\, h(n,\alpha) \geq \min_{0<\alpha<1} c(\alpha) \times \min_{0<\alpha<1} h(n,\alpha)$, we need only find lower bounds for the latter two minima.

Lemma 1: $\min_{0<\alpha<1} c(\alpha) > 2.029$.

Proof: $\min\limits_{0<\alpha<1} c(\alpha) = c(\sqrt{\sqrt{2}-1}) = \sqrt{2+3/\sqrt{2}} > 2.029$. *q.e.d.*

We shall later need the following:

Lemma 2: $\sum\limits_{k=m+1}^{n+1} \frac{1}{k} < \log\left(\frac{n+1}{m}\right) < \sum\limits_{k=m}^{n} \frac{1}{k}$ for $n \geq m \geq 1$.

Proof: By elementary calculus, $\frac{1}{k+1} < \int_k^{k+1} \frac{1}{x}\, dx < \frac{1}{k}$. Thus

$$\sum\limits_{k=m}^{n} \frac{1}{k+1} < \int_m^{n+1} \frac{1}{x}\, dx = \log\left(\frac{n+1}{m}\right) < \sum\limits_{k=m}^{n} \frac{1}{k} \quad \text{for } n, m \geq 1 . \qquad q.e.d.$$

Now we can find a lower bound for $\min\limits_{0<\alpha<1} h(n,\alpha)$.

Lemma 3: $\min\limits_{0<\alpha<1} h(n,\alpha) > n^{\log \sqrt{2}} > n^{0.3465}$.

Proof: Let $w(n) = \sum\limits_{k=2}^{n} 1/(k-1)$. If $p = n$, then

$h(n,\alpha) = (1/\alpha)^{w(n)}$. If $p = 0$, then $h(n,\alpha) = (1/\sqrt{1-\alpha^2})^{w(n)} t(n)$,

where $t(n) = \prod\limits_{j=1}^{n/2} \left[\left(\frac{2j}{2j-1}\right) \left(\frac{2j}{1-\alpha}\right)^{1/(2j-1)} \right]^{1/(2j-2)} \geq 1$.

Thus $h(n,\alpha) \geq b(\alpha)^{w(n)}$, where $b(\alpha) = \max\{1/\alpha, 1/\sqrt{1-\alpha^2}\}$. But

$\min\limits_{0<\alpha<1} b(\alpha) = \sqrt{2}$, and is attained by $\alpha = 1/\sqrt{2}$. From Lemma 2,

$w(n) > \log(n)$. Hence $\min\limits_{0<\alpha<1} h(n,\alpha) > (\sqrt{2})^{\log n} = n^{\log \sqrt{2}} > n^{0.3465}$

q.e.d.

From Lemmas 1 and 3 we obtain a lower bound for $c(\alpha)\, h(n,\alpha)$

for $0 < \alpha < 1$.

Theorem 1: $\min\limits_{0<\alpha<1} c(\alpha)\, h(n,\alpha) > (2.029) n^{0.3465}$

In §12.4 we shall show that $c(\alpha_0) \, h(n,\alpha_0) < 3.07(n-1)^{0.446}$.

Thus our choice of $\alpha = \alpha_0$ (§5.7) will provide us with an upper bound

on $c(\alpha_0) \, h(n,\alpha_0)$ which does not differ much from the minimum of

$c(\alpha) \, h(n,\alpha)$ for $0 < \alpha < 1$.

12.3   Remarks on an Upper Bound for $h(n,\alpha)$

We cannot evaluate the β_k until we know the pivotal selection,

but if we could bound all the β_k independent of the pivotal selection,

then we can obtain an upper bound.

Theorem 2: If $\beta_k \leq c$ for all $k \geq r$, where $c > 0$ and $r \geq 3$

is independent of n, then for $0 < \alpha < 1$:

$$h(n,\alpha)/h(r-1,\alpha) < \gamma(n-1)^{\log \sqrt{c}} \;, \text{ where } \gamma = (r-2)^{-\log \sqrt{c}}$$

Proof: From Lemma 2 of §12.2, $\displaystyle\sum_{k=r\geq 3}^{n} 1/(k-1) < \log \left(\frac{n-1}{r-2} \right)$.

Thus $h(n,\alpha)/h(r-1,\alpha) \leq \{ \displaystyle\prod_{k=r}^{n} c^{1/(k-1)} \}^{1/2} < c^{\frac{1}{2}\log \left(\frac{n-1}{r-2} \right)}$.

But $x^{\log y} = y^{\log x}$ for $x,y > 0$.   *q.e.d.*

This shows that if we can bound β_k independent of the pivotal

strategy for $k \geq r$ and if we can consider the worst possible cases

of pivot selection for $k = 2,\ldots,r-1$ and bound $h(r-1,\alpha)$, then we

have a bound for $h(n,\alpha)$. (In order to consider the cases for

$2 \leq k \leq r-1$, we must have r reasonably small.)

In §12.4 we shall do this for $\alpha = \alpha_0$, and we shall have $r = 5$.

12.4   <u>An</u> <u>A</u> <u>Priori</u> <u>Bound</u> <u>for</u> $\alpha = \alpha_0 = (1 + \sqrt{17})/8$

For $\alpha = \alpha_0$, max $(1/\alpha_0^2, \sqrt{(1 + \alpha_0^2)/(1 - \alpha_0^2)}) = 1/\alpha_0$. Thus, for $\alpha = \alpha_0$,

$$(12.2.1) \quad c(\alpha_0) = d(\alpha_0)/\alpha_0 \text{ , where } d(\alpha) = \begin{cases} 1 & \text{if pivot[1] = 1} \\ \\ 1/\sqrt{1 - \alpha^2} & \text{if pivot[2] = 2} \end{cases}$$

Now we shall show that for $\alpha = \alpha_0$, $\beta_k^{1/(k-1)}$ is maximal for

1×1 pivots for $k \geq 5$. Recall §9.1 or §11.2 for the definition of

pivot[k] .

<u>Lemma</u> 4: For $\alpha = \alpha_0$ and $k \geq 5$, if pivot[k] = 0 , then

$$\beta_k^{1/(k-1)} \beta_{k+1}^{1/k} \leq (1/\alpha_0^2)^{1/(k-1) + 1/k} .$$

<u>Proof</u>: From (11.5.1), if pivot[k] = 0 , then for $0 < \alpha < 1$,

$\beta_{k+1} = 1/(1 - \alpha^2)$ and $\beta_k = \frac{1}{k} [\frac{k+1}{1-\alpha^2}]^{1+1/k}$. Hence

$\beta_k^{1/(k-1)} \beta_{k+1}^{1/k} \leq (1/\alpha^2)^{1/(k-1) + 1/k}$ iff $(k+1)^{k+1}/k^k \leq \alpha^2 (1/\alpha^2 - 1)^{2k}$

iff $(k+1) \log (k+1) - k \log k \leq \log \alpha^2 + 2 k \log (1/\alpha^2 - 1)$

iff $g(k,\alpha) \geq 0$ for $k \geq k_0$ for some integer $k_0 \geq 1$,

where $g(x,\alpha) = 2 x \log (1/\alpha^2 - 1) + \log \alpha^2 - (x+1) \log (x+1) + x \log x$.

$$\frac{\partial}{\partial x} g(x,\alpha) = 2 \log (1/\alpha^2 - 1) - \log (1 + 1/x) > 0$$

iff $x > 1/[(1/\alpha^2 - 1)^2 - 1]$.

In order to evaluate this inequality, we now fix $\alpha = \alpha_0$. Let

$G(x) = g(x, \alpha_0)$.

Thus $G'(x) = \frac{\partial}{\partial x} g(x, \alpha_0) > 0$ iff $x > 0.94$. So $G(x)$ is a monotone increasing function for $x > 0.94$. Now $G(4) = -0.476$ and $G(5) = 3.698$. Thus $g(k, \alpha_0) > 0$ for $k \geq 5$. $q.e.d.$

Now we can apply Theorem 2 of §12.3 and Lemma 4 to obtain the following two necessary lemmas.

<u>Lemma</u> 5: If $pivot[5] \neq 2$, then $h(n, \alpha_0)/h(4, \alpha_0) < 3^{\log \alpha_0} (n-1)^{-\log \alpha_0}$
$< 0.613 (n-1)^{0.446}$.

<u>Proof</u>: By Theorem 2 of §12.2 and Lemma 4, $h(n, \alpha_0)/h(4, \alpha)$
$< 3^{\log \alpha_0} (n-1)^{-\log \alpha_0} < 0.613 (n-1)^{0.446}$. $q.e.d.$

<u>Lemma</u> 6: If $pivot[5] = 2$, then $h(n, \alpha_0)/h(3, \alpha_0) < 0.717 (n-1)^{0.446}$.

<u>Proof</u>: If $pivot[5] = 2$, then $pivot[6] \neq 2$, so by Theorem 2 of §12.2 and Lemma 4, $h(n, \alpha)/h(5, \alpha) < 4^{\log \alpha_0} (n-1)^{-\log \alpha_0}$. Since

$pivot[5] = 2$, $\{\beta_4^{\frac{1}{3}} \beta_5^{\frac{1}{4}}\}^{1/2} = \{[\frac{1}{4} (\frac{5}{1-\alpha_0^2})^{5/4}]^{1/3} (\frac{1}{1-\alpha_0^2})^{1/4}\}^{1/2}$.

$4^{\log \alpha_0} \{\beta_4^{\frac{1}{3}} \beta_5^{\frac{1}{4}}\}^{1/2} < 0.717$. $q.e.d.$

Now we need only to bound $\{\beta_2 \beta_3^{\frac{1}{2}}\}^{1/2}$ for $pivot[5] = 2$ and
$\{\beta_2^{\frac{1}{2}} \beta_3^{\frac{1}{3}} \beta_4\}^{1/2}$ for $pivot[5] \neq 2$.

<u>Lemma</u> 7: If $pivot[3] = 1$, then $h(3, \alpha_0) = \begin{cases} (1/\alpha_0)^{3/2} & \text{if } pivot[2] = 1 \\[2em] \frac{1}{\sqrt{\alpha_0(1-\alpha_0^2)}} & \text{if } pivot[2] = 2 \end{cases}$

< 1.96.

Proof: Since pivot[3] $= 1$, $\beta_3 = \frac{1}{\alpha_0^2}$ and pivot $[2] \neq 0$.

So $\beta_2 = \begin{cases} \frac{1}{\alpha_0^2} & \text{if pivot}[2] = 1 \\ \\ \frac{1}{1-\alpha_0^2} & \text{if pivot}[2] = 2 \end{cases}$, and $h(3,\alpha_0) < 1.96$.

$q.e.d.$

Lemma 8: If pivot[3] $= 2$, then $h(3,\alpha_0) < 2.74$.

Proof: Since pivot[3] $= 2$, $\beta_3 = \frac{1}{1-\alpha_0^2}$ and $\beta_2 = \frac{1}{2}\left(\frac{3}{1-\alpha_0^2}\right)^{3/2}$.

$q.e.d.$

Lemma 9: If pivot[3] $= 0$, then

$$h(4,\alpha_0) = \frac{4^{1/3}}{3^{\frac{1}{4}}\sqrt{1-\alpha_0^2}} \times \begin{cases} \frac{1}{\alpha_0} & \text{if pivot}[2] = 1 \\ \\ \frac{1}{\sqrt{1-\alpha_0^2}} & \text{if pivot}[2] = 2 \end{cases}.$$

Proof: Since pivot[3] $= 0$, $\beta_4 = \frac{1}{1-\alpha_0^2}$, $\beta_3 = \frac{1}{3}\left(\frac{4}{1-\alpha_0^2}\right)^{4/3}$,

and pivot[2] $\neq 0$. So $\beta_2 = \begin{cases} \frac{1}{\alpha_0^2} & \text{if pivot}[2] = 1 \\ \\ \frac{1}{1-\alpha_0^2} & \text{if pivot}[2] = 2 \end{cases}$. $q.e.d.$

Now we put Lemmas 5-9 together to get a bound on $c(\alpha_0)\, h(n,\alpha_0)$ which is independent of the pivotal selection.

Theorem 3: $c(\alpha_0)\,h(n,\alpha_0) < 3.07\,(n-1)^{0.446} < 3\,\sqrt{n}$ for $n \geq 2$.

Proof: From (12.2.1), $c(\alpha_0) = d(\alpha_0)/\alpha_0$,

$$d(\alpha_0) = \begin{cases} 1 & \text{if } pivot[1] = 1 \\ \\ 1/\sqrt{1 - \alpha_0^2} & \text{if } pivot[2] = 2 \end{cases}$$

Now we must consider various situations.

<u>Case</u> I: $pivot[5] = 2$: Then $pivot[3] \neq 0$. By Lemma 6,

$h(n,\alpha_0)/h(3,\alpha_0) < 0.717\,(n-1)^{0.446}$.

(a) If $pivot[3] = 1$:

Then, by Lemma 7, $h(3,\alpha_0) < 1.96$. Since $d(\alpha_0) \leq 1/\sqrt{1 - \alpha_0^2}$,
$c(\alpha_0)\,h(n,\alpha_0) < 2.39\,(n-1)^{0.446}$.

(b) If $pivot[3] = 2$:

Then $pivot[1] = 1$. So $d(\alpha_0) = 1$. By Lemma 8, $h(3,\alpha_0) < 2.74$.
So $c(\alpha_0)\,h(n,\alpha_0) < 3.07\,(n-1)^{0.446}$.

<u>Case</u> II: $pivot[5] \neq 2$: Then $pivot[4] \neq 0$. By Lemma 5,

$h(n,\alpha_0)/h(4,\alpha_0) < 0.613\,(n-1)^{0.446}$.

(a) If $pivot[4] = 2$:

By Lemma 9, $d(\alpha_0)\,h(4,\alpha_0) < 2.58$. Thus $c(\alpha_0)\,h(n,\alpha_0) <$
$2.48\,(n-1)^{0.446}$.

(b) If $pivot[4] = 1$:

Then $\beta_4 = 1/\alpha_0^2$ and $pivot[3] \neq 0$.

(i) If pivot[3] = 2 :

Then pivot[1] = 1 , so $d(\alpha_0) = 1$. By Lemma 8,

$h(3,\alpha_0) < 2.74$. Thus $c(\alpha_0)h(n,\alpha_0) <$

$(1/\alpha_0)(2.74)\ \beta_4^{1/6}\ (0.613)(n-1)^{0.446} < 3.04\ (n-1)^{0.446}$.

(ii) If pivot[3] = 1 :

By Lemma 7, $h(3,\alpha_0) = \begin{cases} (1/\alpha_0)^{3/2} & \text{if pivot[2] = 1} \\[2ex] \dfrac{1}{\sqrt{\alpha_0(1-\alpha_0^2)}} & \text{if pivot[2] = 2} \end{cases}$.

We must use this form of Lemma 7 in order to prove the theorem.
Once again we have two cases.

(A) If pivot[2] = 1 :

Then pivot[1] = 1 , and $d(\alpha_0) = 1$. So

$c(\alpha_0)\ h(n,\alpha_0) < (1/\alpha_0)^{17/6}\ (0.613)(n-1)^{0.446} <$

$2.17\ (n-1)^{0.446}$.

(B) If pivot[2] = 2:

Then $d(\alpha_0) = 1/\sqrt{1-\alpha_0^2}$. So $c(\alpha_0)\ h(n,\alpha_0) <$

$(1/\alpha_0)^{11/6}\ [1/(1-\alpha_0^2)]\ (0.613)(n-1)^{0.446} < 2.36\ (n-1)^{0.446}$.

Thus, in all cases, we have $c(\alpha_0)\ h(n,\alpha_0) < 3.07\ (n-1)^{0.446}$,

for $n \geq 2$.

Now $3.07\ n^{0.446} < 3\sqrt{n}$ for $n \geq 2$.

q.e.d.

12.5   Bound on Element Growth

From §11.7 and §12.5, we see that the elements in all the reduced matrices are bounded by $\sqrt{n} \, f(n) \, \mu_0 \, (3.07)(n-1)^{0.446}$, where $\mu_0 = \max_{i,j} |A_{ij}|$, under strategy S_α with $\alpha = \alpha_0 = (1 + \sqrt{17})/8$ for the pivotal strategies described in §6.1, §6.4, and particularly §8.1.

For Gaussian elimination with complete pivoting (Wilkinson, 1961; pp. 281-285), the bound on element growth is $\sqrt{n} \, f(n) \, \mu_0$.

Thus, for $\alpha = \alpha_0$, our bound for diagonal pivoting is within a factor $3.07(n-1)^{0.446}$ of Wilkinson's bound for Gaussian elimination with complete pivoting.

12.6   Conjecture for Gaussian Elimination

It is conjectured that the best possible bound is n for real matrices under Gaussian elimination with complete pivoting (Cryer, p. 343). The conjecture is false for complex matrices (Tornheim, 1965). For real matrices, the best possible bound is n for $n = 1,2,4$ and is 2-$1/4$ for $n = 3$.

A matrix H is a Hadamard matrix if $|H_{ij}| = 1$ and the rows of H are orthogonal. Then the order of H is 2 or is divisible by 4 (Davis, p. 327). Under Gaussian elimination, $|H^{(1)}| \geq n$ (Cryer, p. 343). Thus, a Hadamard matrix of order n has element growth of at least n .

If H is generated by tensor products of $\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$ (Davis, p. 326),

and if the order of H is $n = 2^k$, then $|H^{(1)}| = n = 2^k$.

12.7 Conjecture for Diagonal Pivoting

If H is as in the previous paragraph, then H is symmetric and
the diagonal pivoting method uses a 1×1 pivot at each step under stra-
tegy S_α for $0 < \alpha < 1$, and $|H^{(1)}| = n = 2^k$.

We conjecture that the optimal bound for diagonal pivoting is of

the form $n\, q(\alpha)$. We need a function $q(\alpha) \geq 1$ since $\begin{bmatrix} -\beta & 1 \\ 1 & \beta \end{bmatrix}$

has $\alpha + 1/\alpha$ as its bound on element growth, where $0 < \alpha \leq \beta \leq 1$.
Thus, for $n = 2$, $q(\alpha) = (\alpha + 1/\alpha)/2$, and $q(\alpha_0) = 1.10$.

Or, we could conjecture a best possible bound of the form
$n\, q(n,\alpha)$. Then $q(2,\alpha) = (\alpha + 1/\alpha)/2$ and $q(2,\alpha_0) = 1.10$.

12.8 The Optimal Choice of α

The optimal choice of α for $0 < \alpha < 1$ is the value which mini-
mizes $q(\alpha)$ in §12.7 for all n . Or, we could seek a sequence of α_n
such that α_n minimizes $q(n,\alpha)$ in §12.7. But we do not know $q(\alpha)$
or $q(n,\alpha)$ for $n > 2$.

We could choose α to minimize $c(\alpha)\, h(n,\alpha)$ in (11.7.11). But
$\sqrt{n}\, f(n)\, \mu_0\, c(\alpha)\, h(n,\alpha)$ is merely a bound on the element growth and is,
by no means, the best possible bound.

Since $c(\alpha_0)\, h(n,\alpha_0) < 3.07\, (n-1)^{0.446}$ (§12.4), while

$$\min_{0<\alpha<1} c(\alpha)\, h(n,\alpha) > (2.029)\, n^{0.3465}$$ (§12.2), our choice of $\alpha = \alpha_0$

gives us a bound (independent of the pivotal selection) for

$c(\alpha_0)\, h(n,\alpha_0)$ which does not differ much from a lower bound for

$$\min_{0<\alpha<1} c(\alpha)\, h(n,\alpha)\,.$$ We assert, further, that $q(\alpha_0)$ is not much greater

than $\min_{0<\alpha<1} q(\alpha)\,.$

We note that $\inf_{0<\alpha<1} q(2,\alpha) = 1 = q(2,1)$. But this implies that

$\alpha_2 = 1$, so we would use a 1×1 iff one of the diagonal elements were

maximal, i.e. if $\mu_1 < \mu_0$ we must use a 2×2 and thus for $n = 2$ no

decomposition would be performed, which is to be expected since the

minimal element growth occurs when we do no decomposition at all. Thus

$\inf_{0<\alpha<1} q(2,\alpha) = 1 = q(2,1)$ does not provide us with any information for

the general n^{th} order case if we require α_n to minimize $q(n,\alpha)$.

Chapter 13 : Iterative Improvement

13.1   The Approximate Solution

Let us assume now that we have obtained an approximate solution z to the system $A x = b$, $A = A^t$, $\det A \neq 0$, by the method of diagonal pivoting. The approximate solution z to $A x = b$ can be considered the exact solution to the system $(A + E) z = b$.

In exact arithmetic, the method of diagonal pivoting would perform the following steps (see §10.2):

(1) $A = M D M^t$

(2) $c = M^{-1} b$

(3) $y = D^{-1} c$

(4) $x = M^{-t} y$

However, in finite precision we have error at each step. For some error matrices F , M_1 , D , and M_2 , we actually perform (see §10.2):

(1) $A + F = M D M^t$

(2) $c = (M + M_1)^{-1} b$

(3) $y = (D + \delta D)^{-1} c$

(4) $x = (M + M_2)^{-t} y$.

From §10.3, we see that z is the exact solution to $(A + E)z = b$, where

$$E = F + M_1 (D + \delta D)(M + M_2)^t + M [\delta D (M + M_2)^t + D M_2^t] .$$

From (10.15.6), we have for $\alpha = \alpha_0$ ($\|\ \|$ is the one-norm):

$$\|E\| < (23.54) \; n^2 \; 2^{-t} \; \max_{k} \; \mu_0^{(k)} \; .$$

In Chapters 11 and 12 we have shown for $\alpha = \alpha_0$:

$$\max_{k} \; \mu_0^{(k)} < \sqrt{n} \; f(n) \; \mu_0 \; (3.07)(n-1)^{0.446} \; ,$$

where $\mu_0 = \max_{i,j} \; |A_{ij}|$. Hence, for $\alpha = \alpha_0$:

$$\|E\| < (72.3) \; n^{2.9446} \; f(n) \; \mu_0 \; 2^{-t} \; .$$

13.2   The Iteration

Let $x_1 = z$. For $m = 1,2,\ldots$ we obtain an improved solution x_{m+1} to $A \, x = b$ by the following

(1) $r_m = b - A \, x_m$

(2) $(A + E) \, d_m = r_m$

(3) $x_{m+1} = x_m + d_m$.

We shall assume that (1) and (3) are done exactly. This is a reasonable assumption if accumulated inner products are used (see Wilkinson (1965), pp. 116-117).

Thus we shall assume the only error occurred when we tried to solve $A \, d_m = r_m$ but performed (2) $(A + E) \, d_m = r_m$ instead.

The iteration is meaningful only if the matrix A can be represented exactly in the computer, i.e. if the elements of A are known exactly and no round-off occurs when the numbers are read into the computer.

13.3   Convergence of the Iteration

We must now show that the x_m defined in §13.2 converge to the solution $x = A^{-1} b$.

Theorem: Let $x = A^{-1} b$. Assume $\|A^{-1} E\| = \sigma < 1/2$. Let $r_j = b - A x_j$, $(A + E) d_j = r_j$, and $x_{j+1} = x_j + d_j$. Then $\lim_{m \to \infty} \|x_m - x\| = 0$.

Proof: $(A + E) d_{m-1} = r_{m-1} = b - A x_{m-1}$. So $(A + E)(x_m - x_{m-1}) = A x - A x_{m-1}$. Thus $(A + E)(x_m - x) = E(x_{m-1} - x)$. Since $\det A \neq 0$, $(I + A^{-1} E)(x_m - x) = A^{-1} E(x_m - x)$.

Let $\sigma = \|A^{-1} E\| < 1/2$. Let $\tau = \sigma/(1 - \sigma)$. So $\tau < 1$. $\|x_m - x\| \leq \|x_{m-1} - x\| \, \|A^{-1} E\| / (1 - \|A^{-1} E\|) = \tau \|x_{m-1} - x\|$

$\leq \tau^{m-1} \|x_1 - x\|$. Since $\tau < 1$, $\lim_{m \to \infty} \|x_m - x\| = 0$.   *q.e.d.*

Thus the iterative vectors converge to the solution $x = A^{-1} b$ provided that $\|A^{-1} E\| < 1/2$. and the convergence in monotone in the norm, i.e. $\|x_{m+1} - x\| \leq \|x_m - x\|$.

Now we must give conditions under which $\|A^{-1} E\| < 1/2$. We shall use the one-norm: $\|x\| = \sum_{i=1}^{n} |x_i|$ and $\|A\| = \max_{j} \sum_{i=1}^{n} |A_{ij}|$.

Corollary: If $(72.3) \, n^{2.9446} \, f(n) \, \kappa(A) \, 2^{-t} < 1/2$, where $\kappa(A) = \|A\| \, \|A^{-1}\|$, then $\lim_{m \to \infty} \|x_m - x\| = 0$ where $x = A^{-1} b$.

Proof: From §13.1, $\|E\| < (72.3) \, n^{2.9446} \, f(n) \, \mu_0 \, 2^{-t}$. But $\mu_0 \leq \|A\|$, so $\|A^{-1}\| \, \mu_0 \leq K(A)$. Hence $\|A^{-1} \, E\| < (72.3) \, n^{2.9446}$ $f(n) \, \kappa(A) \, 2^{-t}$, and the corollary follows from the previous theorem.

$q.e.d.$

We see that the convergence of the iterates depends on the condition number of A , and so we cannot expect iterative improvement to be of value for ill-conditioned matrices. For further remarks on condition numbers, see Wilkinson (1965); and on iterative improvement, see Fox (pp. 49-53, 109-113) and Moler (pp. 316-321). As we noted in §13.2 the iteration is meaningful only if no round-off occurred while reading A into the computer and if the elements of A are exact.

Chapter 14 : Symmetric Band Matrices

14.1   Gaussian Elimination for Band Matrices

Let A be an n×n non-singular band matrix with band widrh
$2m+1 \ll n$, i.e. $A_{ij} = 0$ if $|i-j| > m$.

We could store A in $(2m+1)n$ locations rather than n^2 locations by ignoring A_{ij} for $|i-j| > m$. If we could preserve the band structure while solving $A x = b$, then we would save storage and thus be able to solve band matrices of very large order in relatively few storage locations.

If we use Gaussian elimination with complete pivoting, then we must interchange to bring the maximal element to the leading diagonal position. This could destroy the band structure and we would need n^2 locations to store L and U in the decomposition.

If we use Gaussian elimination with partial pivoting, then L is unit lower triangular with $L_{ij} = 0$ if $|i-j| > m+1$ and U is upper triangular with $U_{ij} = 0$ if $|i-j| > 2m+1$. Thus L can be stored in mn locations and U in $(2m+1)n$ locations. Since L and part of U can be written over A , we would need only mn additional storage locations.

Thus if A is an n×n band matrix with band width $2m+1$ and if b is a vector of length n , then Gaussian elimination with partial pivoting requires only $(3m+2)n$ storage locations to solve $A x = b$. Furthermore, this method requires only $\sim (2m^2 + 4m + 1)n$ multiplications and additions.

14.2 Diagonal Pivoting for Symmetric Band Matrices

Let A be an $n \times n$ symmetric non-singular band matrix with band width $2m+1$. If we use diagonal pivoting (see Chapters 5, 6 and 8) to solve $A x = b$, then interchanges can destroy the band structure and we would need $\frac{1}{2} n^2$ storage locations.

We have investigated many variations of the diagonal pivoting method for the symmetric band case, but these algorithms have either been unstable or have required more storage and operations than Gaussian elimination with partial pivoting.

At the present time we recommend Gaussian elimination with partial pivoting rather than the diagonal pivoting method (see Chapters 5 and 8) for symmetric band matrices, and thus we are unable to take advantage of the symmetry. (For the special case of symmetric tridiagonal matrices, see §14.3.)

14.3 Symmetric Tridiagonal Matrices

We are able to present a stable algorithm for the symmetric tridiagonal case which requires less storage than does Gaussian elimination with partial pivoting.

Let A be an $n \times n$ symmetric non-singular tridiagonal matrix, i.e. $A_{ij} = 0$ if $|i-j| > 1$. Let $A_{ii} = a_i$ for $1 \le i \le n$ and $A_{i,i+1} = b_i = A_{i+1,i}$ for $1 \le i \le n-1$. Assume $|a_1| \le \alpha$, while $\max_{2 \le i \le n} |a_i|$, $\max_{1 \le i \le n-1} |b_i| \le \beta$.

Let us consider only the first step, which is typical (cf. Appendix A, §A.3).

If $b_1 = 0$, we have nothing to do, and $A_{ij}^{(n-1)} = A_{i+1,j+1}$.

Suppose $b_1 \neq 0$. Let r be a non-negative integer such that $2^r > \beta$.

If $|a_1| \geq 2^{-r} b_1^2$, then $a_1 \neq 0$ and we shall use a_1 as a 1×1 pivot. Then $M_{21} = b_1/a_1$, $A_{11}^{(n-1)} = a_2 - M_{21} b_1$, and $A_{ij}^{(n-1)} = A_{i+1,j+1}$. Thus $A^{(n-1)}$ is tridiagonal, $|M_{21}| \leq 2^r/|b_1|$, $|A_{11}^{(n-1)}| \leq \beta + 2^r$, while $|A_{ij}^{(n-1)}| \leq \beta$ otherwise.

If $|a_1| < 2^{-r} b_1^2$, then we shall use $\begin{bmatrix} a_1 & b_1 \\ b_1 & a_2 \end{bmatrix}$ as a 2×2 pivot. (Note that we make no interchanges.) Then

$$|a_1 a_2 - b_1^2| \geq b_1^2 - |a_1 a_2| > b_1^2 (1 - 2^{-r} |a_2|) \geq b_1^2 (1 - 2^{-r} \beta) > 0$$

since $2^{-r} \beta < 1$ by assumption.

Now $M_{31} = -b_1 b_2/(a_1 a_2 - b_1^2)$, $M_{32} = a_1 b_2/(a_1 a_2 - b_1^2)$, $A_{11}^{(n-2)} = a_3 - M_{32} b_2$, and $A_{ij}^{(n-2)} = A_{i+2,j+2}$ otherwise. Thus $A^{(n-2)}$ is tridiagonal, $|M_{31}| \leq |b_2|/[|b_1| (1 - 2^{-r} \beta)]$, $|M_{32}| \leq 2^{-r} \beta/(1 - 2^{-r} \beta)$, $|A_{11}^{(n-2)}| \leq \beta/(1 - 2^{-r} \beta)$, while $|A_{ij}^{(n-2)}| \leq \beta$ otherwise.

We see that the bounds on the elements of $A^{(n-1)}$ for a 1×1 and $A^{(n-2)}$ for 2×2 are independent of the bound α on $|a_1|$.

Thus the pattern continues throughout all the reduced matrices. Hence we conclude : each reduced matrix $A^{(k)}$ is tridiagonal,

$$|A_{11}^{(k)}| \leq \max \{\beta + 2^r , \beta/(1 - 2^{-r} \beta)\} , \text{ while } |A_{ij}^{(k)}| \leq \beta$$

otherwise.

Usually we normalize by choosing $\beta = 1$. Then,

$$|A_{11}^{(k)}| \leq \max \{1 + 2^r , 1/(1 - 2^{-r})\} , \text{ while } |A_{ij}^{(k)}| \leq 1 \text{ otherwise,}$$

for each reduced matrix $A^{(k)}$. Since $2^r > \beta = 1$, $r \geq 1$. Hence

$$|A_{11}^{(k)}| \leq 1 + 2^r . \text{ Thus given any positive integer } r , \text{ we have}$$

$$\max_{k} \max_{i,j} |A_{ij}^{(k)}| \leq 1 + 2^r \quad (= 3 \text{ for } r = 1) .$$

A backward error analysis of this algorithm shows that it is very stable (since the elements of all the reduced matrices are bounded by $1 + 2^r$, which takes on its minimal value 3 for $r = 1$).

Thus we can decompose $A = M D M^t$, where D is block diagonal with blocks of order 1 and 2, and M is unit lower triangular with $M_{i+1,i} = 0$ if $D_{i+1,i} \neq 0$ and with $|M_{ij}| = 0$ if $i > j+2$.

We shall need an n-vector array to record the pivotal selection. We set pivot[k] = 1 (2) if we use a 1×1 (2×2) pivot for $A^{(k)}$. If pivot[k] = 2 , we set pivot[k-1] = $M_{n-k+3,n-k+1}$. Then we need only 2n storage locations to store the rest of M and D (these we write over A). Thus we need only 3n storage locations for this algorithm.

From §14.2, we see that Gaussian elimination requires 5n storage and 7n operations, and is very stable for tridiagonals (see Appendix A, §A.3).

We would also like the number of operations required for our algo-- rithm to be less than $7n$. However, if we use the algorithm in the manner in which we have expressed it, $8\frac{1}{2} n - 2\frac{1}{2} p$ multiplications and $5n$ additions are required, where p is the number of 1×1 pivots used. Thus between $6n$ and $8\frac{1}{2}$ multiplications are required. (We ignore multiplication by 2^r in the count.)

However, we can reduce the number of multiplications from $8\frac{1}{2} n - 2\frac{1}{2} p$ to $7\frac{1}{2} n - \frac{3}{2} p$ if we implement the algorithm in the following manner:

(We present the first step of the algorithm in Algol form):

```
if b[1] = 0 then M[2,1] := 0

  else  begin temp := a[1]/b[1];

          if abs (temp) >= abs (b[1])×2↑(-r) then

            begin M[2,1] := 1/temp; a[2] := a[2] - M[2,1]×b[1] end

          else  begin calc := temp×a[2] - b[1];

                  M[3,1] := - b[2]/calc; M[3,2] := - temp×M[3,1];

                  a[3] := a[3] - M[3,2]×b[2] end

      end;
```

A backward error analysis shows that this implementation of the algorithm is also very stable (since all the reduced matrices are bounded by $1 + 2^r$).

Now $\min_{r\geq1} (1 + 2^r) = 3$ for $r = 1$, while $(7\frac{1}{2} n - \frac{3}{2} p) \to 6n$ as $r \to \infty$ (since the larger r is the more likely the choice of a

1×1 pivot becomes). But $(1 + 2^r) \to \infty$ as $r \to \infty$, and thus we would not have a good bound on the error matrix for large r. Thus in practice, we must make some reasonable choice of $r \geq 1$ so that $1 + 2^r$ is not too large but so that $7\frac{1}{2} n - \frac{3}{2} p$ is reasonably small (i.e. as close to $6n$ as possible, and hopefully not more than $7n$).

We have considered many versions of diagonal pivoting for the tridiagonal case. The minimal storage possible is $3n$. The above-mentioned algorithm had the least operation count of all the versions studied.

Since this algorithm requires between $6n$ and $7\frac{1}{2} n$ multiplications in comparison to $7n$ for Gaussian elimination with partial pivoting, we can recommend this algorithm for general use only if storage of $3n$ rather than $5n$ is crucial to the user.

Appendix A : Miscellaneous Results

A.1   Diagonal Pivoting for Positive Definite Matrices

If A is an n×n symmetric positive definite matrix, then the maximal element of A is on its diagonal. So $\mu_0 = \mu_1$ and, according to S_α , we use that maximal diagonal element as pivot. But $A^{(n-1)}$ is also positive definite. Thus p = n (where p is the number of 1×1 pivots used in the decomposition.)

Since $\mu_0^{(k)} = \mu_1^{(k)}$ for each $A^{(k)}$, calculating $\mu_0^{(k)}$ is unnecessary. (This calculation would require $\sim \frac{1}{6} n^3$ additions for all the $\mu_0^{(k)}$.) Thus if we know that A is positive definite, we may omit the calculation of the $\mu_0^{(k)}$, and our method is identical to the method of congruent transformations (§2.7). If we also omit the calculation of $\mu_1^{(k)}$ and use the first diagonal element as a 1×1 pivot (non-zero since A is positive definite), our method is identical to L D L^t (§2.7).

From the above we see that the L D L^t method and the method of congruent transformations (i.e. L D L^t with pivoting on the diagonal) are special cases of the diagonal pivoting method, and either of these may be used if A is definite. (See §§2.6 - 2.11 for further remarks on this topic.)

In our algorithm for the diagonal pivoting method in Appendix C, we allow the following options:

(1) If A is indefinite, then we must use diagonal pivoting.

(2) If A is definite, then we may use:

(a) $L D L^t$ with pivoting on the diagonal (by omitting the calculation of the $\mu_0^{(k)}$), or

(b) $L D L^t$ (by omitting the calculation of the $\mu_0^{(k)}$ and the $\mu_1^{(k)}$).

A.2   A Result for Symmetric Hadamard Matrices

An n×n real matrix H is Hadamard if $|H_{ij}| = \mu_0$ for all i,j where $\mu_0 > 0$, and $H H^t = n \mu_0^2 I$. Usually we normalize by choosing $\mu_0 = 1$. Thus all the elements of H are of the same modulus, and the columns of H re mutually orthogonal, i.e. if H_j is the jth column of H , then $H_i^t H_j = n \delta_{ij}$ where

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ & \qquad \text{is the Kronecker delta.} \\ 0 & \text{if } i \neq j \end{cases}$$

If the diagonal pivoting method or symmetric Gaussian elimination is used to solve H x = b , then H_{11} is used as the pivot for the first step under any strategy. Then the reduced matrix $H^{(n-1)}$ has the following interesting properties:

Theorem: $H^{(n-1)}$ is not Hadamard, and the angle between any two columns of $H^{(n-1)}$ is $\pi/3$.

Proof: Let us assume $\mu_0 = 1$. We use H_{11} as pivot. Let $B = H^{(n-1)}$.

Then $B_{ij} = H_{i+1,j+1} - H_{i+1,1} H_{j+1,1}/H_{11}$ for $1 \leq i,j \leq n-1$.

$$B_r^t B_s = \sum_{k=1}^{n-1} B_{rk} B_{ks} = \sum_{k=1}^{n-1} (H_{r+1,k+1} - H_{r+1,1} H_{k+1,1}/H_{11}) \times$$

$$(H_{s+1,k+1} - H_{s+1,1} H_{k+1,1}/H_{11})$$

$$= H_{r+1}^t H_{s+1} - H_1^t H_{s+1} (H_{r+1,1}/H_{11}) - H_1^t H_{r+1}(H_{s+1,1}/H_{11}) +$$

$$H_1^t H_1 (H_{r+1,1} H_{s+1,1}/H_{11}^2) = H_{r+1}^t H_{s+1} + H_1^t H_1 (H_{r+1,1} H_{s+1,1}/H_{11}^2) .$$

If $r \neq s$, then $H_{r+1}^t H_{s+1} = 0$, so $B_r^t B_s = \pm n$ since $H_{ij} = \pm 1$ for all i,j and $H_1^t H_1 = n$.

Thus $B_r^t B_s \neq 0$ for $r \neq s$, so $H^{(n-1)}$ is not Hadamard.

Further $B_r^t B_r = n + n (H_{r+1,1}^2/H_{11}^2) = 2n$. Define $\|B_r\|^2 = B_r^t B_r$ and $\cos O(r,s) = B_r^t B_s/(\|B_r\| \|B_s\|)$. Then

$\cos O(r,s) = \pm 1/2$, so $O(r,s) = \pm \pi/3$ for $r \neq s$.

$q.e.d.$

A similar result holds for Hermitian Hadamard matrices.

A.3   Gaussian Elimination for Tridiagonals

Let T be an $n \times n$ tridiagonal matrix (i.e. $T_{ij} = 0$ for $|i-j| > 1$). Suppose $|T_{11}| \leq \alpha$, and $|T_{ij}| \leq \beta$ otherwise.

Theorem: Let T be as above. Then for any reduced matrix $T^{(k)}$ under Gaussian elimination with partial pivoting : $T^{(k)}$ is tridiagonal, $|T_{11}^{(k)}| \leq 2\beta$, and $|T_{ij}^{(k)}| \leq \beta$ otherwise.

Proof: The situation for $k = n-1$ is typical.

If $|T_{11}| \geq |T_{21}|$, then we use T_{11} as the pivot. So $T_{11}^{(n-1)} = T_{22} - T_{21} \, _{12}/T_{11}$, while $T_{ij}^{(n-1)} = T_{i+1,j+1}$ otherwise. Thus $T^{(n-1)}$ is tridiagonal, $|T_{11}^{(n-1)}| \leq 2\beta$, and $|T_{ij}^{(n-1)}| \leq \beta$ otherwise.

If $|T_{11}| < |T_{21}|$, then we interchange the first and second rows and use T_{21} as the pivot. So $T_{11}^{(n-1)} = T_{12} - T_{11} T_{22}/T_{21}$, $T_{12}^{(n-1)} = -T_{23} T_{11}/T_{21}$, while $T_{ij}^{(n-1)} = T_{i+1,j+1}$ otherwise. Thus $T^{(n-1)}$ is tridiagonal, $|T_{11}^{(n-1)}| < 2\beta$, and $|T_{ij}^{(n-1)}| \leq \beta$ otherwise. $q.e.d.$

From the theorem and §2.5, we conclude that Gaussian elimination with partial pivoting is very stable for tridiagonal matrices T , since $\max\limits_{k} \max\limits_{i,j} |T_{ij}^{(k)}| \leq 2 \max\limits_{i,j} |T_{ij}|$.

Appendix B : Algorithm for Symmetric Equilibration

B.1   Discussion

The following Algol procedure will equilibrated any $n \times n$ symmetric matrix A so that $D A D$ is equilibrated, where D is diagonal. A is replaced by $D A D$, and the inverses of the diagonal elements of D are stored in the vector d. (See Chapter 7.)

B.2   The Algol Procedure

```
procedure symequil (A, n, d);
   value n; array A,d; integer n;
   comment the symmetric matrix  A  of order  n  is equili-
      brated and the symmetric equilibrated matrix  D A D  is
      stored in A , where  D⁻¹ = diag (d[1], ..., d[n]);
   begin integer i,j; real t;
      for i := 1 step 1 until n do
        begin d[i] := sqrt (abs (A[i,i]));
           for j := 1 step 1 until i-1 do
              begin t := abs (A[i,j]);
                 if t > d[i] then d[i] := t
              end;
           if d[i] ≠ 0 then
              begin for j := 1 step 1 until i do
                       A[i,j] := A[i,j]/d[i];
                    for j := i step 1 until n do
                       A[j,i] := A[j,i]/d[i];
```

```
            end;

         end;

      for i := 1 step 1 until n do

         if d[i] = 0 then

            begin  for j := i+1 step 1 until n do

                      begin t := abs (A[j,i]);

                         if t > d[i] then d[i] := t

                      end;

                   if d[i] = 0 then goto alarm;

                   for j := i+1 step 1 until n do

                      A[j,i] := A[j,i]/d[i];

                   goto out;

            end;

         alarm: print ('this matrix has a null row')

out : end;
```

Appendix C : <u>Algorithm for Diagonal Pivoting</u>

The following listing of an Algol procedure will solve $A X = B$ by the diagonal pivoting method, where A is an $n \times n$ non-singular symmetric matrix and B is a vector of length n .

The $L D L^t$ method (symmetric Gaussian elimination) and the method of congruent transformations $(L D L^t$ with pivoting on the diagonal) are special cases of the diagonal pivoting algorithm.

The matrix A is assumed to be stored only in its lower triangular part. A is decomposed into $A = M D M^t$, where M is unit lower triangular, D is symmetric block diagonal with blocks of order 1 or 2, and $M[i+1,i] = 0$ when $D[i+1,i] \neq 0$. M and D are written over the lower triangular part of A .

A is declared $[1 : n, 1 : n]$ and B is $[1 : n]$. Upon exit, the solution X to $A X = B$ is stored in B , i.e., $X[i]$ is stored in $B[i]$.

If A is indefinite, then set DEF $= 0$ and the general diagonal pivoting method is used.

If A is (positive or negative) definite, then we may omit the calculation of the maximum off-diagonal element in the reduced matrices. If DEF $= 2$ then this is omitted and the algorithm is identical to $L D L^t$ with pivoting on the diagonal. The pivoting on the diagonal

may also be omitted if desired by setting   DEF = 1 , and then the algorithm is identical to L D L^t .

The algorithm,as presented below, is by no means, in its most efficient form.   In particular, as written, no advantage of symmetry is taken to reduce storage.   Instead of using only the lower triangular part of A[1:n, 1:n], the algorithm should be coded so that the lower triangular part of A is stored in a one-dimensional array of length $\frac{1}{2}$ n (n+1) .   Further, B[1 : n] could be replaced by B[1:n, 1:k] for solving a system with k right hand sides.

```
#PROCEDURE# PIVOTING (A,B,N,DEF) ..
   #VALUE# N,DEF ..
   #ARRAY# A,B .,      #INTEGER# N,DEF ..

   #COMMENT# SOLVES A X = B BY THE DIAGONAL PIVOTING METHOD
            WHERE A IS A SYMMETRIC MATRIX OF ORDER N AND
            B IS A VECTOR OF LENGTH N ..
   #COMMENT# A IS ASSUMED TO BE STORED ONLY IN ITS LOWER
            TRIANGULAR PART, M AND D ARE WRITTEN OVER A WHERE
            A = M D M TRANSPOSE, M IS UNIT LOWER TRIANGULAR,
            AND D IS BLOCK DIAGONAL WITH BLOCKS OF ORDER 1 OR 2,
            AND M(/I+1,I/) = 0 WHEN D(/I+1,I/) #NOT EQUAL# 0 ..
   #COMMENT# IF A IS INDEFINITE, SET DEF = 0 AND THE DIAGONAL
            PIVOTING METHOD IS USED..
   #COMMENT# IF A IS (POSITIVE OR NEGATIVE) DEFINITE, THEN
            SET DEF = 1 AND L D L TRANSPOSE WITHOUT PIVOTING
            WILL BE USED, OR SET DEF = 2 AND L D L TRANSPOSE
            WITH PIVOTING ON THE DIAGONAL WILL BE USED ..

   #BEGIN# #INTEGER# I,J,K,R,S ..
#REAL# M0, M1, DET, SAVE, TEMP, ALPHA ..
   #INTEGER# #ARRAY# CHANGE (/1..N/) ..
#ARRAY# PIVOT (/1..N/) ..

#PROCEDURE# MAXDIAG (A,K,N,J,M1) ..
   #VALUE# K,N .. #ARRAY# A .. #INTEGER# K,N,J ..
        #REAL# M1 ..
#COMMENT# CALCULATES THE MAXIMUM OF THE DIAGONAL OF A , M1 =
   MAX ABS(A(/I,I/)) FOR K #LEQ# I #LEQ# N, AND J IS THE
   LEAST INTEGER SUCH THAT M1 = ABS(A(/J,J/)) ..
#BEGIN# #INTEGER# I .. M1 .= ABS(A(/K,K/)) .. J .= K ..
   #FOR# I .= K+1 #STEP# 1 #UNTIL# N #DO#
        #IF# ABS(A(/I,I/)) #GREATER# M1 #THEN#
           #BEGIN# M1 .= ABS(A(/I,I/)) .. J .= I #END# ..
#END# MAXDIAG ..

   #PROCEDURE# MAX A (A,K,N,R,S,M0,L,M1) ..
      #VALUE# K,N,L,M1 ..
      #ARRAY# A .. #INTEGER# K,N,R,S,L .. #REAL# M0,M1 ..
        #COMMENT# CALCULATES M0 = MAX ABS(A(/I,J/) ) FOR
            L #LEQ# I,J #LEQ# N AND THE INTEGERS R AND S SUCH
            THAT ABS(A(/R,S/)) = M0, IT IS ASSUMED THAT M1 =
            MAX (ABS ( A(/I,I/) ) ) AND M1 = ABS( A(/L,L/) ) ..
        #BEGIN# #INTEGER# I,J .. M0 .= M1 .. R .= S .= L ..
        #FOR# J .= K #STEP# 1 #UNTIL# N-1 #DO#
```

```
      #FOR# J .= K #STEP# 1 #UNTIL# N #DO#
      #FOR# I .= J+1 #STEP# 1 #UNTIL# N #DO#
          #IF# AHS ( A(/I.J/) ) #GREATER# MO #THEN#
              #BEGIN# MO .= AHS ( A(/I.J/) ) .+
                      K .= I .. S .= J #END# .+
    #END# MAXO .+

  #PROCEDURE# INTERCHANGE (A.K.I) .+
    #VALUE# K.I .. #ARRAY# A .. #INTEGER# K.I ..
    #COMMENT# INTERCHANGES ROW AND COLUMN K WITH ROW
        AND COLUMN I WHERE K #GEQ# I AND A IS THE REDUCED
        MATRIX OF ORDER N-I+1 .+
    #BEGIN# #REAL# TEMP .. #INTEGER# J ..
      #FOR# J .= K+1 #STEP# 1 #UNTIL# N #DO#
          #BEGIN# TEMP .= A(/J.K/) .. A(/J.K/) .= A(/J.I/) ..
              A(/J.I/) .= TEMP #END# ..
      #FOR# J .= I+1 #STEP# 1 #UNTIL# K-1 #DO#
          #BEGIN# TEMP .= A(/J.I/) .. A(/J.I/) .= A(/K.J/) ..
              A(/K.J/) .= TEMP #END# ..
      TEMP .= A(/I.I/) .. A(/I.I/) .= A(/K.K/) ..
      A(/K.K/) .= TEMP
    #END# INTERCHANGE ..

ALPHA .= (1 + SQRT(17) )/8 .+
    I .= 1 ..

START..

    #IF# DEF = 1 #THEN#
        #BEGIN# CHANGE (/I/) .= I .. #GOTO# PIVOTONE #END# .+
    MAXDIAG (A.I.N.K.M1) ..
    #IF# DEF = 2 #THEN#
        #BEGIN# INTERCHANGE (A.K.I) .. CHANGE (/I/) .= K ..
        #GOTO# PIVOTONE #END# .+
    MAXO (A.I.N.R.S.MO.K.M1) ..
    #IF# M1 #NOT LESS# MO + ALPHA #THEN#
    #BEGIN# INTERCHANGE (A.K.I) .. CHANGE (/I/) .= K ..
        #GOTO# PIVOTONE #END# .+
        CHANGE (/I/) .= S .+
    #IF# S #GREATER# I #THEN#  INTERCHANGE (A.S.I) ..
        CHANGE (/I+1/) .= R .+
    #IF# R #GREATER# I+1 #THEN# INTERCHANGE (A.R.I+1) ..
#GOTO# PIVOTTWO .+

PIVOTONE ..

    #IF# A(/I.I/) = 0 #THEN# #GOTO# ALARM .+
#COMMENT#   WE USE A 1X1 PIVOT ..
    #FOR# J .= I+1 #STEP# 1 #UNTIL# N #DO#
      A(/J.I/) .= A(/J.I/)/ A(/I.I/) ..
        #COMMENT# A(/J.I/) HAS BEEN SET EQUAL TO THE MULTIPLIER ..
    #FOR# J .= I+1 #STEP# 1 #UNTIL# N #DO#
    #FOR# K .= I+1 #STEP# 1 #UNTIL# J #DO#
      A(/J.K/) .= A(/J.K/) - A(/J.I/)*A(/K.I/) ..
```

```
        #COMMENT# THE A(/J,K/) HAVE BEEN SET TO THEIR NEW VALUES ..
#COMMENT# PIVOT(/I/) = 1 IF WE USE A 1X1 AT ROW I ..
    PIVOT (/I/) .= 1 ..    I .= I+1 ..
    #IF# I #NOT GREATER# N #THEN# #GOTO# START
          #ELSE# #GOTO# FIND C ..


PIVOTTWO ..

        DET .= A(/I,I/)*A(/I+1,I+1/) - A(/I+1,I/)*A(/I+1,I/) ..
    #IF# DET = C #THEN# #GOTO# ALARM ..
#COMMENT#   WE USE A 2X2 PIVOT ..
    #FOR# J .= I+2 #STEP# 1 #UNTIL# N #DO#
    #BEGIN#    #FOR# K .= I+2 #STEP# 1 #UNTIL# J-1 #DO#
          A(/J,K/) .= A(/J,K/) - A(/K,I/)*A(/J,I/) - A(/K,I+1/)*
                     A(/J,I+1/) ..
    #COMMENT# THE A(/J,K/) HAVE BEEN SET TO THEIR NEW VALUES ..
          SAVE .= A(/J,I/) .. TEMP .= A(/J,I+1/) ..
    A(/J,I/) .= (A(/I+1,I+1/)*SAVE - A(/I+1,I/)*TEMP)/DET ..
    A(/J,I+1/) .= (A(/I,I/)*TEMP - A(/I+1,I/)*SAVE)/DET ..
    #COMMENT# A(/J,I/) AND A(/J,I+1/) HAVE BEEN SET EQUAL
              TO THE APPROPRIATE MULTIPLIER ..
    A(/J,J/) .= A(/J,J/) - A(/J,I/)*SAVE - A(/J,I+1/)*TEMP ..
    #END# ..
#COMMENT# PIVOT(/I/) = 2 IF WE USE A 2X2 AT ROW I AND THEN DET IS
          STORED IN PIVOT(/I+1/) ..
    PIVOT (/I/) .= 2 .. PIVOT (/I+1/) .= DET .. I .= I+2 ..
    #IF# I #NOT GREATER# N #THEN# #GOTO# START
          #ELSE# #GOTO# FIND C ..


#COMMENT# NOW FORM C = M INVERSE TIMES B AND STORE IT IN B ..

FIND C ..

    I .= 1 ..

REPEAT..

        SAVE .= B(/I/) .. B(/I/) .= B(/CHANGE(/I/) /) ..
        B(/ CHANGE(/I/) /) .= SAVE ..
    #IF# PIVOT(/I/) = 1 #THEN#    #BEGIN#
    #FOR# J .= I+1 #STEP# 1 #UNTIL# N #DO#
        B(/J/) .= B(/J/) - A(/J,I/) * B(/I/) ..
    I .= I+1 #END#    #ELSE#    #BEGIN#
        SAVE .= B(/I+1/) .. B(/I+1/) .= B(/ CHANGE(/I+1/) /) ..
        B(/ CHANGE(/I+1/) /) .= SAVE ..
    #FOR# J .= I+2 #STEP# 1 #UNTIL# N #DO#
        B(/J/) .= B(/J/) - A(/J,I/)*B(/I/)
                         - A(/J,I+1/)*B(/I+1/) ..
    I .= I+2 #END# ..
    #IF# I #NOT GREATER# N #THEN# #GOTO# REPEAT ..

    I .= 1 ..

#COMMENT# NOW SOLVE  L Y = C  AND STORE Y IN THE VECTOR B ..
```

```
SOLVE ..

    #IF# PIVOT (/I/) = 1 #THEN#
      #BEGIN# B(/I/) .= B(/I/) / A(/I,I/) ..
        I .= I+1 ..
        #IF# I #GREATER# N #THEN# #GOTO# FIND X
          #ELSE# #GOTO# SOLVE ..
      #END# ..
    TEMP .= B(/I/) .. SAVE .= B(/I+1/) .. DET .= PIVOT(/I+1/) ..

    B(/I/) .= ( TEMP*A(/I+1,I+1/) - SAVE*A(/I+1,I/) )/DET ..
    B(/I+1/) .= ( SAVE*A(/I,I/) - TEMP*A(/I+1,I/) )/DET ..
    I .= I+2 ..
    #IF# I #GREATER# N #THEN# #GOTO# FINDX #ELSE# #GOTO# SOLVE ..

#COMMENT#   NOW SOLVE  X = M INVERSE TRANSPOSE TIMES Y WHERE Y IS
       STORED IN THE VECTOR B AND STORE X IN B ..

FINDX ..

    I .= N ..
    CALC ..  #IF# PIVOT(/I/) = 1 #THEN#
      #BEGIN#
        #FOR# J .= I+1 #STEP# 1 #UNTIL# N #DO#
        B(/I/) .= B(/I/) - A(/J,I/)*B(/J/) ..
        SAVE .= B(/I/) .. B(/I/) .= B(/ CHANGE(/I/) /) ..
        B(/ CHANGE(/I/) /) .= SAVE ..   I .= I-1 #END#
      #ELSE# #BEGIN#   #FOR# K .= I-1,I #DO#
        #BEGIN#
        #FOR# J .= I+1 #STEP# 1 #UNTIL# N #DO#
          B(/K/) .= B(/K/) - A(/J,K/)*B(/J/) ..
      #END# ..     #FOR# K .= I-1, I #DO#      #BEGIN#
          SAVE .= B(/K/) .. B(/K/) .= B(/ CHANGE(/K/) /) ..
          B(/ CHANGE(/K/) /) .= SAVE .. #END#..
          I .= I-2 #END# ..
    #IF# I #NOT LESS# 1 #THEN# #GOTO# CALC ..

#GOTO# OUT ..

ALARM ..

    OUTPUT (61, #(#   #(# SINGULAR MATRIX #)#    #)#   ) ..

OUT .. #END# PIVOTING ..
```

Bibliography

Cryer, C. W., "Pivot size in Gaussian elimination", _Numerische_
  Mathematik, 12 (1968), pp. 335-345.

Davis, P. J., _The Mathematics of Matrices_, Blaisdell, New York (1965).

Forsythe, G. and C. Moler, _Computer Solution of Linear Algebraic_
  Systems, Prentice-Hall, New Jersey (1967).

Fox, L., _An Introduction to Numerical Linear Algebra_, Oxford Uni-
  versity Press, London (1964).

Gantmacher, F. R., _The Theory of Matrices_, v. 1, Chelsea, New York
  (1959).

Hildebrand, F. B., _Introduction to Numerical Analysis_, McGraw-Hill,
  New York (1956).

Householder, A. S., _The Theory of Matrices in Numerical Analysis_,
  Blaisdell, New York (1964).

Marcus, M. and M. Newman, "The permanent of a symmetric matrix",
  Notices Amer. Math. Soc., v. 8 (1961), p. 595.

Meersman, R. de and L. Schotsmans, "Note on the inversion of symmetric
  matrices by the Gauss-Jordan method", _I.C.C. Bulletin_, 3 (1964),
  pp. 152-5.

Mirsky, L., _An Introduction to Linear Algebra_, Clarendon Press,
  Oxford (1955).

Moler, C., "Iterative refinement in floating point", _Journal A.C.M._,
  14:2 (1967), pp. 316-321.

Parlett, R. N., and J. K. Reid, "On the solution of a system of
    linear equations whose matrix is symmetric but not definite",
    B.I.T. (to appear).

Reid, J. K., "A note on the least square solution of a band system
    of linear equations by Householder reductions", Computer Journal
    10:2 (1967), pp. 188-189.

Schotsmans, L., "Gauss-Jordan inversion of symmetric matrices, using
    a 2×2 pivot submatrix", C.E.N, Report No. 621-42/65-153, Mol,
    Belgium (1965).

Shisha, O., editor, Inequalities, Academic Press, New York (1967).

Sinkhorn, R., and P. Knopp, "Concerning non-negative matrices and
    doubly stochastic matrices", Pacific Journal of Math., 21:2
    (1967).

Sinkhorn, R., "Diagonal equivalence to matrices with prescribed row
    and column sums", Amer. Math. Monthly, 74:4 (1967).

Sinkhorn, R., "A relationship between arbitrary positive matrices and
    doubly stochastic matrices", Ann. Math. Statistics, 35 (1964),
    pp. 876-879.

Tornheim, L., "Pivot size in Gauss reduction", Tech. Report, Calif.
    Res. Corp., Richmond, Calif. (1964).

Tornheim, L., "Maximum third pivot for Gaussian reduction", Tech.
    Report, Calif. Res. Corp., Richmond, Calif. (1965).

Westlake, J., Handbook of Numerical Matrix Inversion and Solution of
    Linear Equations, Wiley, New York (1968).

Wilkinson, J. H., "Rounding errors in algebraic processes", Infor-
    mation Processing, "Proceedings of the International Conference
    on Information Processing, UNESCO, Paris (1959).

Wilkinson, J. H., "Error analysis of direct methods of matrix inversion", _Journal A.C.M._, 8:3 (1961), pp. 281-330.

Wilkinson, J. H., _Rounding errors in algebraic processes, Notes on Applied Science No. 32_, Her Majesty's Stationery Office, London; Prentice-Hall, New Jersey (1963).

Wilkinson, J. H., _The Algebraic Eigenvalue Problem_, Clarendon Press, Oxford (1965).

END